ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

A study of Chinese learners' prepositional bias in Spanish from a cognitive-linguistic perspective.

Dexin Kong[1,*]

[1] Xi'an Fanyi University, Xi'an, Shaanxi, 710000, China

SUMMARY: *For second language learners, language learning is not simply a matter of vocabulary building, especially when it comes to the richness and variety of prepositions, which are subject to bias. This study investigates Chinese learners' misuse of Spanish prepositions, explains the transfer between native cognitive schemas and target language structures from a cognitive linguistic perspective, and provides a new approach to understanding the cognitive mechanisms of second language acquisition. A statistical linguistics-based text retrieval model (SLM-IR), including probabilistic model and language generation model, is introduced and optimized. On the other hand, rooted in the prototype and schema theories of cognitive linguistics, it is argued that the bias arises from the learner's attempt to understand and apply the Spanish prepositional system with the conceptual categories inherent in the native language. Experiments confirmed that a statistical language model using Bayesian smoothing techniques performed best in distinguishing prepositional positives and negatives, with a word perplexity of 472 and a MAP = 75.34%. Meanwhile, the TF-IDF weighting strategy achieves an optimal balance between both model size retrieval performance, with a model size of only 6.96 megatrigram when the word perplexity is 500, achieving a MAP of 75.15%. The usage of Spanish prepositions by Chinese learners, influenced by the Chinese word "de", overuses "de", with its usage frequency accounting for as high as 27.47%. The error rates of "en" and "para" are prominent, being 12.99% and 15.77% respectively. The alternative category of bias accounts for 61.89% of the total bias, suggesting that the learner difficulty lies in the misalignment of conceptual mappings.*

KEYWORDS: *cognitive linguistics; Spanish prepositions; usage bias; statistical linguistics; schema theory*

1 Introduction

Prepositions play a key role in the grammatical system as specific words or affixes whose main function is to identify the grammatical relations and semantic roles of the immediately following words (usually pronouns or phrases with nominal properties) in a sentence [1-4]. Prepositions are often placed before these words, and together they build prepositional phrases that convey a rich variety of contextual information such as time, place, reason, purpose, manner, status, or comparative relations [5-7]. If the preposition is removed from the original sentence, it will lead to a significant change in the meaning of the sentence. Chinese prepositions play an important role in many sentences, and they are also a key and challenging knowledge in the Chinese grammatical system [8-10]. Spanish is characterized by inflectional language, rich morphological changes in words and close syntactic relationships [11]. The two

*Kmiranda@163.com
https://doi.org/10.65102/is2026099
Published: 30 April, 2026

languages, Chinese and Spanish, are quite different from each other, and their prepositions differ in their expressions [12, 13]. In the cognitive linguistics perspective, due to the complexity of preposition usage, Chinese learners are susceptible to the negative transfer effect of their mother tongue in the process of acquiring Spanish, which can easily lead to prepositional errors [14-17].

The analysis of bias has a positive significance for Chinese learners' acquisition of Spanish, so in teaching practice, students' bias is summarized in a timely manner, and the bias should be guided and corrected in a targeted manner [18-20]. According to the different learning stages of students, the knowledge of complex morphological changes in Spanish should be summarized and organized, and more effective language input materials should be designed and compiled for teaching [21, 22]. In teaching, we can also use more Chinese-Spanish comparisons to help them eliminate the interference of their mother tongue, overcome the phenomenon of "rigidity", and establish an intermediary language system similar to the rules of Spanish at an early date [23-25].

Traditionally, the analysis of bias in the use of prepositions by second language learners has focused on the comparison of grammatical forms. Based on the cognitive linguistics perspective, the study penetrates into the learner's cognitive world to attribute the causes of the bias. First, based on two text retrieval models, the desired language examples are found from the massive text materials. The probabilistic model (PM), based on Bayes' theorem, transforms language retrieval into a probabilistic inference problem based on word occurrence. The difference in the probability of occurrence of words in relevant and non-relevant sets is utilized to assess the document value. Statistical language model-based retrieval is elevated from whether a document contains a query word to what kind of topic probability distribution the document as a whole presents. The unification of frequency statistics (TF) and discriminative statistics (IDF) in a probabilistic framework is revealed. Smoothing is also introduced to compensate for smoothing with a global language model of the entire document set. It also explores how probabilistic models estimate relevance and how concepts such as frequency and usage define the activity of a language component. Finally, a perspective on the application of cognitive linguistics in language teaching is introduced. Cognitive prototype theory and schema theory are elaborated to help understand the reasons for learners' biases in foreign language use.

2 Overview of Cognitive Linguistics-based Statistical Language Processing Models and Word Statistics Research

2.1 Text retrieval model

2.1.1 Probabilistic Model (PM)

Probabilistic models are based on the theory that given a user's query string q and a document d^j in a collection, a probabilistic model to estimate the probability that the user's query string is related to the document d^j. The probabilistic model assumes that this probability is determined only by the query string and the documents. Further, the model assumes that there exists a set of all documents, i.e., a subset of resultant documents with respect to the query string q, and that such an ideal set is denoted by R, and that the documents in the set are expected to be relevant to the query string. The definition for computing the probability of relevance is given below.

The weights of index terms in probabilistic models are binary, e.g., $w_{k,j} \in \{0,1\}$, $w_{k,q} \in \{0,1\}$. The query string q is a subset of the set of index terms. Let R be the set of relevant documents (the initial set of guesses) and $\bar{R}$ be the complement of R (the set of non-relevant documents). $P\left(R\left|\vec{d}_j\right.\right)$ denotes the probability that document d^j is related to query string q, and $P\left(\bar{R}\left|\vec{d}_j\right.\right)$ denotes the probability that document d^j is not related to the query string q. The relevance value of a document d^j for a query string q is defined as $sim\left(d_j,q\right)=\frac{P\left(R\left|\vec{d}_j\right.\right)}{P\left(\bar{R}\left|\vec{d}_j\right.\right)}$, according to Bayesian's law:

$$sim\left(d_j,q\right)=\frac{p\left(\vec{d}_j\left|R\right.\right)\times p\left(R\right)}{p\left(\vec{d}_j\left|\bar{R}\right.\right)\times p\left(\bar{R}\right)} \tag{1}$$

$p\left(\vec{d}_j\left|R\right.\right)$ represents the probability of randomly selecting a document d^j from the set of related documents R. $p\left(R\right)$ represents the probability of randomly selecting a document from the entire set as the relevant document. Similarly define $p\left(\vec{d}_j\left|\bar{R}\right.\right)$, $p\left(\bar{R}\right)$. Since $p\left(R\right)$ and $p\left(\bar{R}\right)$ are the same for all documents in the set, the

$$\mathrm{sim}\left(d_j,q\right)\propto\frac{p\left(\vec{d}_j\left|R\right.\right)}{p\left(\vec{d}_j\left|\bar{R}\right.\right)} \tag{2}$$

Assuming that the indexing terms are independent of each other then:

$$sim\left(d_j,q\right)\propto\frac{\prod_{k\in(t_j,1)}p\left(k_i\left|R\right.\right)\times\prod_{k\in(t_j,0)}p\left(\bar{k}_i\left|R\right.\right)}{\prod_{k\in(t_j,1)}p\left(k_i\left|\bar{R}\right.\right)\times\prod_{k\in(t_j,0)}p\left(\bar{k}_i\left|\bar{R}\right.\right)} \tag{3}$$

$p\left(k_i\left|R\right.\right)$ denotes the probability that an indexing term k_i occurs in a randomly selected document in set R, and $p\left(\bar{k}_i\left|R\right.\right)$ denotes the probability that an indexing term k_i does not occur in a randomly selected document in set R, similarly defined $p\left(k_i\left|\bar{R}\right.\right)$, $p\left(\bar{k}_i\left|\bar{R}\right.\right)$. Taking logarithms, according to $p\left(k_i\left|R\right.\right)+p\left(\bar{k}_i\left|R\right.\right)=1$, is finally obtained:

$$sim\left(d_j,q\right)\propto\sum_{i=1}^{I}w_{i,q}\times w_{i,j}\times\left(\log\frac{p\left(k_i\left|R\right.\right)}{1-p\left(k_i\left|R\right.\right)}+\log\frac{p\left(k_i\left|\bar{R}\right.\right)}{1-p\left(k_i\left|\bar{R}\right.\right)}\right) \tag{4}$$

This is a key expression for computing correlation in a probabilistic model. Since the set R is not known at the beginning, an algorithm must be devised that initializes the computation of $p\left(k_i \middle| R\right)$ and $p(k_i \mid \bar{R})$.

At the beginning of the query only the query string is defined and the set of resultant documents is not yet available.

Assumption (a) assumes that $p\left(k_i \middle| R\right)$ is constant (generally equal to 0.5) for all index terms k_i;

Assumption (b) assumes that the distribution of index terms in unrelated documents can be approximated by the distribution of index terms in all documents in the set.

These two assumptions are expressed in the following formulas:

$$p\left(k_i \middle| R\right) = 0.5 \tag{5}$$

$$p\left(\bar{k}_i \middle| R\right) = \frac{n_i}{N} \tag{6}$$

n_i denotes the number of documents in which the index term k_i occurs, and N is the number of total documents in the set. Under the above assumptions, we can obtain documents that partially contain the query string and provide them with an initial probability of relevance.

The advantage of the probabilistic model is that the documents can be ranked in order of their decreasing probability of relevance. His disadvantage lies in the fact that at the beginning it is necessary to guess the division of the documents into two sets, relevant and irrelevant, and in fact this model does not take into account the frequency of the index terms in the documents (since all the weights are binary), which are independent of each other.

2.1.2 Statistical Linguistic (SLM) Modeling Based IR Modeling

Investigate the application of statistical language modeling to the IR domain for text retrieval. The study of topic-based language modeling expresses the prior probability $P_s(t)$ of a document-dependent phrase t as $P_d(t) = \sum_z p(t \mid z)\, p(z \mid d)$. It means: for document d, select topic z with probability $P(z \mid d)$, and then generate lexical entries t by topic z with probability $P(t \mid z)$.

In this paper, the two models of the information retrieval process are defined as shown in Fig. 1.

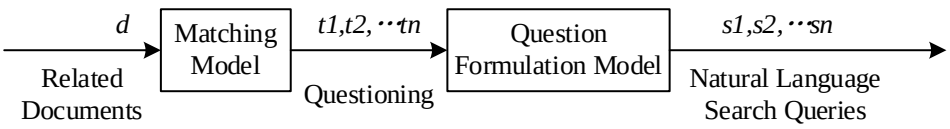


Figure 1: Two models in the information retrieval process

The related document d is "corrupted" into a question $t_1, \cdots, t_n$ when it passes through the first noise channel, and then the question is "corrupted" into a request when it passes through the second noise channel. The question is then "corrupted" into a request when it passes through the second noisy channel $s_1, \cdots, s_n$. The information retrieval system can be thought of as a

decoding function $f: s_1 \cdots s_n \to d$ which tries to reproduce the originally sent message, i.e., to find the document related to the request.

The research goal is to find such a decoding function $f\left(s_1 \cdots s_n\right)$ that makes:

$$f\left(s_1 \cdots s_n\right) = \arg\max_d P\left(D = d \mid S_1 = s_1, \cdots, S_n = s_n\right) \tag{7}$$

According to Bayes' formula, there are:

$$P\left(S_1 = s_1, \cdots, S_n = s_n, D = d\right) = \frac{P\left(D = d \mid S_1 = s_1, \cdots, S_n = s_n\right)}{P\left(S_1 = s_1, \cdots, S_n = s_n\right)} \tag{8}$$

Since $P\left(S_1 = s_1, \cdots, S_n = s_n\right)$ is independent of d, Eq. (7) can be simplified to:

$$\begin{aligned} f\left(s_1 ... s_n\right) &= \arg\max_d P\left(S_1 = s_1, ..., S_n = s_n, D = d\right) \\ &= \arg\max_d \sum_{t_1 ... t_n} P\left(S_1 = s_1, ..., S_n = s_n, T_1 = t_1, ..., T_n = t_n, D = d\right) \end{aligned} \tag{9}$$

Since there are two independent channels, equation (9) can be written in the following form:

$$\begin{aligned} f\left(s_1 \cdots s_n\right) &= \arg\max_d P\left(S_1 = s_1, ..., S_n = s_n, D = d\right) \\ &= \arg\max_d \sum_{t_1 ... t_n} P\left(S_1 = s_1, ..., S_n = s_n \mid T_1 = t_1, ..., T_n = t_n\right) \\ &\cdot P\left(T_1 = t_1, ..., T_n = t_n \mid D = d\right) \cdot P\left(D = d\right) \end{aligned} \tag{10}$$

In Eq. (10), $P\left(D = d\right)$ is the a priori probability that document d is relevant; $P\left(T_1 = t_1, ..., T_n = t_n \mid D = d\right)$ is the probability of asking a question given the relevant document; in summary, these two equations define the matching model in Figure 1. $P\left(S_1 = s_1, ..., S_n = s_n \mid T_1 = t_1, ..., T_n = t_n\right)$ is the probability of obtaining a natural language search utterance given the question asked, which defines the question-formulation model in Figure 1.

$P\left(T_1 = t_1, ..., T_n = t_n \mid D = d\right)$ is the language model, and by considering the information retrieval problem through this simple mechanism, the SLM-IR model emerges, as shown in Figure 2.

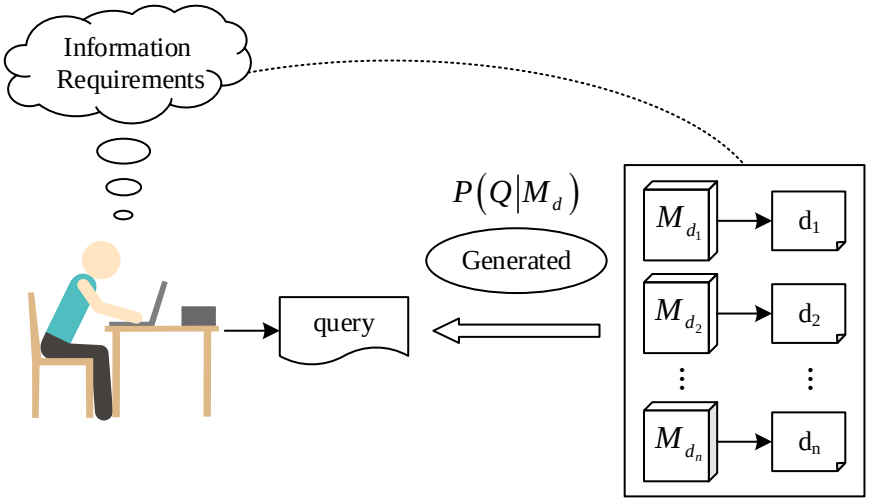


Figure 2: SLM-IR Model

On the surface, the use of a language model may seem very different from a vector space model using a TF-IDF weighting scheme, since the unigram language model already explicitly encodes word frequency (TF), and it seems that the inverted filed frequency (IDF) is not used at all in that model. However, there is an interesting connection between the exploration of language models and traditional models. Much of that connection has to do with smoothing, and a proper evaluation of it can contribute to a deeper understanding of language modeling approaches.

Most smoothing methods use two distributions, model $p_s(w|d)$ for "visible" words appearing in the document, and model $p_u(w|d)$ for "invisible" words. Based on these models, the probability of asking a question q can be written in the following form, where $c(w;d)$ denotes the number of times the word w appears in d.

$$\begin{aligned}\log p(q|d) &= \sum_i \log p(q_i|d) \\ &= \sum_{i:c(q_i;d)>0} \log p_s(q_i|d) + \sum_{i:c(q_i;d)=0} \log p_u(q_i|d) \\ &= \sum_{i:c(q_i;d)>0} \log p_s(q_i|d) - \sum_{i:c(q_i;d)>0} \log p_u(q_i|d) \\ &\quad + \sum_{i:c(q_i;d)>0} \log p_u(q_i|d) + \sum_{i:c(q_i;d)=0} \log p_u(q_i|d) \\ &= \sum_{i:c(q_i;d)>0} \log\left[p_s(q_i|d)/p_u(q_i|d)\right] + \sum_i \log p_u(q_i|d)\end{aligned} \tag{11}$$

The probability of an "unseen" word is usually proportional to the global probability of the word. For example, this is calculated using the document set. Thus, we assume that $p_u(q_i|d) = \alpha_d p(q_i|C)$, where α_d is a document-dependent constant, and $p(q_i|C)$ is the document set language model. From this we get:

$$\begin{aligned}\log p(q|d) &= \sum_{i:c(q_i;d)>0} \log\left[p_s(q_i|d)/\alpha_d p(q_i|C)\right] \\ &\quad + n\log\alpha_d + \sum_i \log p(q_i|C)\end{aligned} \tag{12}$$

Here n is the length of the question. Note that the last item on the right is irrelevant to the document d and can therefore be ignored in the sort.

Now you can see that the search function is actually broken into two parts. The weight of the matching lemma q_i can be measured by $p_s\left(q_i|d\right)/\alpha_d p\left(q_i|C\right)$, which directly corresponds to the word frequency of the document, but conversely to the frequency of the document set.

In this way, the use of $p\left(q_i|C\right)$ as a reference for smoothing the distribution is very similar to the well-known role of IDF. The other part of the formula simply relies on the product of the document's constant and the question length. We can think of it as acting as a document length normalizer, which is an important technique for improving performance in traditional models. In fact, α_d must be closely related to the document length, since we want longer documents to require as little smoothing as possible, and the value of α_d to take a small value. Thus, due to the term, a long document incurs a heavier penalty compared to a short document.

The connection just obtained suggests the use of the document set language model as a reference model for smoothing the document model, implying a retrieval formula that implements the TF-IDF weighting strategy and normalizes the document length. It also shows that smoothing plays a key role in the approach of combining language models and IR.

2.2 Basic Concepts of Word Statistics

In this section, we will explore the basic laws of word occurrence and their statistical representations in the corpus from the perspective of statistical linguistics, so as to provide a methodological basis for the subsequent research on prepositions.

2.2.1 Necessary and random events

In language, there are some rules that are completely deterministic, and if a certain set of deterministic conditions is realized, a completely deterministic event corresponding to it is bound to occur, and such an event is called a necessary event. Abstractly speaking, if a certain definite condition S is realized, a perfectly definite event A corresponding to it must occur, called a necessary event. In communicative activities, this kind of inevitable event is rarely seen. A large number of events in communicative activities do not have such complete certainty, there are usually exceptions.

Because in communicative activities, when a certain group of conditions is realized, the emergence of language components in the vast majority of occasions is not certain, there are exceptions. For example, a certain consonant or rhyme may or may not appear within a certain corpus, and they are random events. Therefore, we can say that the appearance of linguistic components in communicative activities is a random event.

2.2.2 Frequency and probability

In stochastic events, there is a statistical connection between the random event A and the conditional set S, although there is no completely definite connection between them. Although the event A may or may not occur when the conditional set S is realized once, if the conditional set is realized many times, the occurrence of the event A has a certain regularity, which is manifested in the frequency of occurrence of the event A.

The so-called frequency, that is, the actual number of times the event A and the number of times the conditional group S realized ratio, can be expressed in the following formula

$$f = n / N \tag{13}$$

where f denotes the frequency, n is the number of actual occurrences of the event A, and N is the total number of realizations of the conditional group S.

When the number of times the conditional group S is realized is small, the frequency of the random event A is unstable, sometimes it occurs more frequently, sometimes it occurs less frequently, but when the conditional group S is realized many times, with the increase of the number of times it is realized, the frequency of the random event is more and more stable in a definite value, this kind of when the conditional group S is realized many times, the frequency of the random event is becoming more and more stable. The regularity of the condition set S is different from the completely deterministic regularity described earlier, which we call statistical regularity.

Although the appearance of linguistic components in communicative activities is a random event, we can use statistical regularity to characterize it. As long as we reveal this statistical regularity in language, we have a better, more realistic and objective way of describing the various random events in language.

Only by realizing the conditional group S many times is it possible to establish statistical rules for the random event A. Let the number of realizations of the conditional set S be t, as t increases, the frequency of occurrence of the random event A, f, becomes more and more stable, and f converges to a fixed value when $t \to \infty$. This value is called the probability of the random event A, which is denoted as p, i.e.

$$\lim_{t\to\infty} f = \lim_{t\to\infty} \frac{n}{N} = P \tag{14}$$

From the formula, it follows that the probability of a random event is always positive and always lies between 0 and 1, i.e., $0 \le p \le 1$, since n is not greater than N.

If $p = 0$, the random event is an improbable event.

If $p = 1$, the random event becomes completely certain, i.e., a necessary event. It can be seen that a necessary event is nothing but a special case of a random event when $p = 1$. Therefore, in communicative activities, the appearance of language components can be regarded as random events no matter whether they are completely certain or not. If it is a completely determined event, then it can be seen as a special case of random event when $p = 1$.

According to the statistics of 750,000 words in political, literary, scientific and technological publications, it is found that there exists a relationship between the probability of occurrence of a word and its place of occurrence as shown in Figure 3.

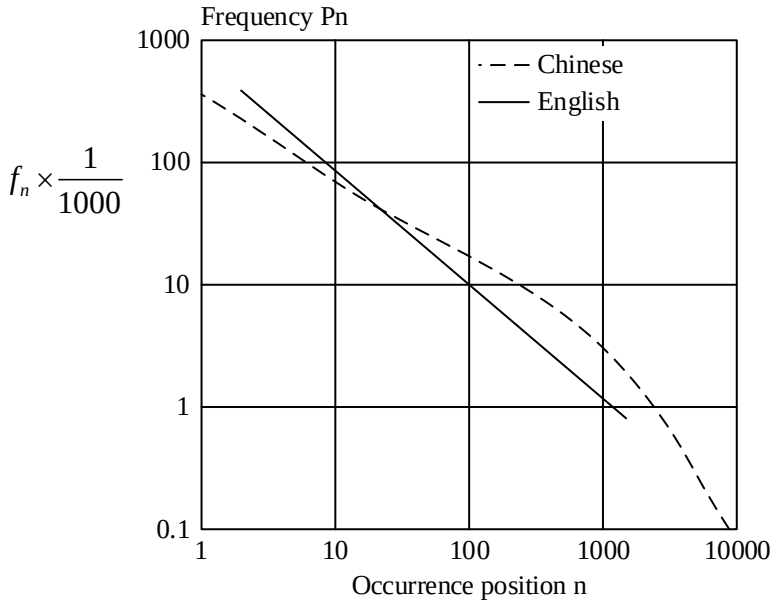


Figure 3: The relationship of the probability of a word's occurrence and its position

The relationship can be written as the following analytic equation:
(1) The word occurs in places between 1 and 100.

$$p_n = 0.033n^{-0.67} \tag{15}$$

(2) Words occurring in places between 100 and 1000.

$$p_n = 0.16n^{-0.96} \tag{16}$$

(3) Words occurring in bits between 1000 and 10000.

$$p_n = 9.0019n^{-0.34} \tag{17}$$

where p_n is the probability of occurrence of the word and n is the word's place of occurrence.

2.2.3 Sampling and sample size

When a linguist studies the facts of a language, he cannot examine the whole totality of the language's texts; he can only take a portion of that totality and study it. This part of the material taken from the totality of the language is called a sample. In this way, the face of the totality can be discussed on the basis of the sample.

Sampling cannot be done arbitrarily, and we must make certain requirements for sampling in order to make our statistical results reliable.

The reliability of statistical results can be expressed in terms of error. Errors can be categorized into absolute and relative errors.

If the sample frequency is denoted as f^* and the overall frequency is denoted as f, the absolute value of the difference between the overall frequency and the sample frequency, $\left| f - f^* \right|$, is called the absolute error. It shows the difference between the sample frequency and

the overall frequency. In order to illustrate the sample frequency and the overall frequency of the degree of proximity, a more appropriate approach is to use the relative error.

The ratio of the absolute error $\left|f - f^*\right|$ to the overall error f is called the relative error and is expressed as δ, i.e.

$$\delta = \left|f - f^* / f\right| \tag{18}$$

Relative errors give a better indication of the precision of the observations than absolute errors. Therefore, it is best to use relative error to account for the reliability of our statistical results.

How exactly should sampling be done to ensure the reliability of the statistical community. The following conditions are proposed for this:

(1) The larger the sample size, the smaller the relative error.

The number of linguistic components contained in the sample is called the sample capacity, denoted as N. When the sample capacity is small the conditional set of random events S is not realized many times and the frequency of the sample is not stable enough. If we observe a large number of texts and sentences, so that the sample capacity is getting bigger and bigger, and thus the number of realizations of the conditional set of random events S increases, then the sample frequency becomes more and more stable. When $N \to 0$, its absolute error $\left|f - f^*\right| \to 0$, and hence its relative error, the

$$\delta = \left|f - f^* / f\right| \to 0 \tag{19}$$

(2) Under a certain relative error δ, the smaller the frequency f, the larger the sample size N is required.

When δ is certain, the smaller the frequency f is, the larger the sample capacity N should be, and when $f \to 0$, then $N \to 0$. This relationship can be expressed by the following formula:

$$\delta = \frac{Z\rho}{\sqrt{Nf}} \tag{20}$$

where δ denotes the relative error, N denotes the sample size, f denotes the frequency, and Zp is a constant, in the general case $Zp = 2$.

2.2.4 Degree of utilization

Usage degree is the compressed word order calculated by a certain formula by combining the three factors of word order, class and article. From this value, we can see the degree of use and distribution of the word in the corpus. The closer the degree of use is to the number of words, the more even the distribution of the number of times the word is used, indicating that the word is more widely used, otherwise the opposite is true. It is closely related to the number of occurrences of the word, but not equivalent. To use the degree of use as a criterion to measure the degree of common use of words is more accurate, reasonable and objective than simply taking the number of times as a criterion. Because, there are often a lot of words, but in favor of a category and a very small number of articles; on the contrary, there are also a small number of words, but the distribution of categories and articles is relatively wide. The term “degree of

usage" is used to show the difference, and the degree of usage is calculated by the corresponding formula for comparing the commonness of words with similar values. Therefore, the top 5,000 words with the highest degree of usage were selected for the objective linguistic probability of the words to be counted.

2.3 Cognitive Linguistics in Foreign Language Teaching and Learning

After clarifying the statistical features of language use, this section returns to the cognitive linguistics perspective to explore how cognitive theories can be applied to language teaching practice, especially in foreign language teaching for Chinese learners. Through the introduction of prototype theory and schema theory, the cognitive mechanisms of learners in language comprehension and use are explained.

2.3.1 Prototype Theory in Cognitive Linguistics

Cognitive linguistics is an emerging fringe discipline formed by the intersection and interpenetration of cognitive science and language science, and its research object is the relationship between cognition and language. The relationship between cognitive science and language science is the basic and primary problem of cognitive linguistics. Cognitive scientists, no matter what kind of specific scientific research they are engaged in, must move towards a point of convergence, i.e., to study language as the most important mental representation, in order to make progress in research. It follows that the relationship between cognition and language is crucial.

Prototype theory can be considered quite important in cognitive linguistics. In layman's terms, when we recognize things, we don't have to have a name for every thing we recognize, which would put a great burden on our memory. When we recognize things, we usually define a category of things. For example, we usually define a cup as a container for drinking water. When we make such a definition, we have ignored the characteristics of each cup, whether it is big, small, wide-mouthed or small, tall or short; instead, we have adopted their common characteristic: a container that can be used for drinking. This is the theory of prototypes.

A prototype is a generalized representation of all the individuals of a category or sphere, reflecting the basic characteristics that a class of objects has. Prototypes give us a framework for understanding the world, facilitating a more rational understanding of the world. And the explanation of the prototype for language lies in the fact that it embodies the economic principles of cognitive behavior and language use. With such a theory as prototype, we can get the maximum cognitive benefit with less cognitive cost when we memorize a large number of things. According to the approach of the prototype theory, the categories of linguistics are not mainly determined by the specificity of the language polytopes, but by human cognition.

2.3.2 Schema Theory in Cognitive Linguistics

With regard to the cognitive-linguistic idea of schema, it is emphasized that human beings rely on structured cognitive frameworks, i.e., "schemas", in order to understand the world. Schemas are organized knowledge structures formed by people based on their previous experiences, which are used to interpret new information, guide behavior and construct meaning. In language comprehension, schemas help listeners or readers activate relevant background knowledge and fill in the implicit information in linguistic expressions, thus realizing efficient communication. For example, in understanding the expression "go to a restaurant", people will automatically activate the "restaurant schema", which includes a series of scripted knowledge such as ordering, eating, paying, etc. In language teaching, the use of schema theory can help learners establish the connection between language and cognitive structure, which is especially important for the

understanding and accuracy of functional items such as prepositions. In language teaching, the use of schema theory can help learners establish the connection between language and cognitive structure, and enhance their spontaneity and accuracy in language processing, especially for the understanding and use of prepositions and other functional lexical items.

3 Statistical Linguistic Modeling and Linguistic Rule Convergence Efficacy Testing

This chapter focuses on the empirical validation of a statistical linguistic model (SLM-IR) and a lexical analysis system that incorporates linguistic rules. It first focuses on the retrieval model itself, verifying the ability of different smoothing techniques to distinguish between correct and incorrect sentences for model prepositions and exploring the balance between model compression and performance with the TF-IDF weighting strategy.Section 3.2 takes a deeper look inside the words, examining the improvement of lexical analysis accuracy when linguistic rules such as lexicality are incorporated into the statistical model.

The study conducts experiments in the Corpus of Learners of Spanish (CORELE), with a sub-collection of Chinese learners as the experimental data source. The sub-corpus is an extensive collection of written Spanish texts produced by Chinese learners of different levels and task types. The corpus was processed with lexical annotation and basic grammatical error labeling.

3.1 Validation of the retrieval model based on statistical language modeling

The experiment will extract all sentences containing the target prepositions from them and classify them into two categories, correct use and biased use, based on manual labeling.

3.1.1 Comparative results of different smoothing techniques

Firstly, we verify the ability of different smoothing techniques to distinguish learners' prepositional usage bias under statistical language model (SLM)-based information retrieval. The SLM-IR model used in this paper is based on the Bayesian smoothing process, selecting the most basic plus-one smoothing (i.e., all word frequencies are added by 1), Goode-Turing estimation (redistributing the probability mass of low-frequency words). Linear interpolation smoothing (linearly weighted mixture of the document language model with the whole collection language model) three smoothing treatments are used as a comparison for bias retrieval.

Given a query containing a target preposition, the system needs to retrieve and sort relevant sentences from a database with a mixture of correct and biased sentences.

Standard information retrieval evaluation metrics are used, including word perplexity, Mean Average Precision (MAP), P@5 (accuracy of the first 5 results), and Normalized Discounted Cumulative Gain (nDCG). These metrics combine to assess the overall quality of the ranked list and the accuracy of the top results.

All of them are based on the statistical language model, and the retrieval performance under the four smoothing processing methods is shown in Table 1.

Table 1: The retrieval performance under four smoothing processing methods

	Word perplexity	MAP/%	P@5/%	nDCG/%
Laplace Smoothing	865	42.83	32.74	0.465
Good-Turing	564	63.11	54.02	0.576
Jelinek-Mercer	491	71.07	63.15	0.727
Dirichlet Prior	472	75.34	67.49	0.786

In this paper, the model under the Bayesian smoothing-based processing method shows stronger applicability, integrating the information of the passage with the global knowledge of the whole corpus. The model perplexity is 472, i.e., the language model with Bayesian smoothing processing is applied to predict the next word, which can be selected with only 472 words on average. Meanwhile, its average accuracy MAP reaches 75.34%, P@5=67.49%, and the normalized discount cumulative gain nDCG = 0.786. It dynamically adjusts the degree of reliance on global information according to document length and content richness.

In contrast, the underlying +1 smoothed word solid width of up to 865 means that the model has to hesitate on average from more than 800 options in predicting each word, with a high level of uncertainty. Its MAP is 42.83%, P@5 32.74%, and only about one-third of the top five results it considers most relevant are indeed correct; The Goode-Turing estimation pass redistributes the probability of low-frequency words and reduces the perplexity to 564. Correspondingly, its various retrieval indexes are also significantly improved, MAP=63.11%, P@5=54.02%; linear interpolation smoothing treatment further strengthens the judgment of the model by linearly weighting, the word perplexity is 491, MAP=71.07%, P@5=63.15%, which is not as good as the paper's Although not as good as the Bayesian smoothing treatment in this paper, its prediction accuracy is still substantially improved by 68.29% compared with the basic plus one smoothing.

3.1.2 Model size effects under different strategies

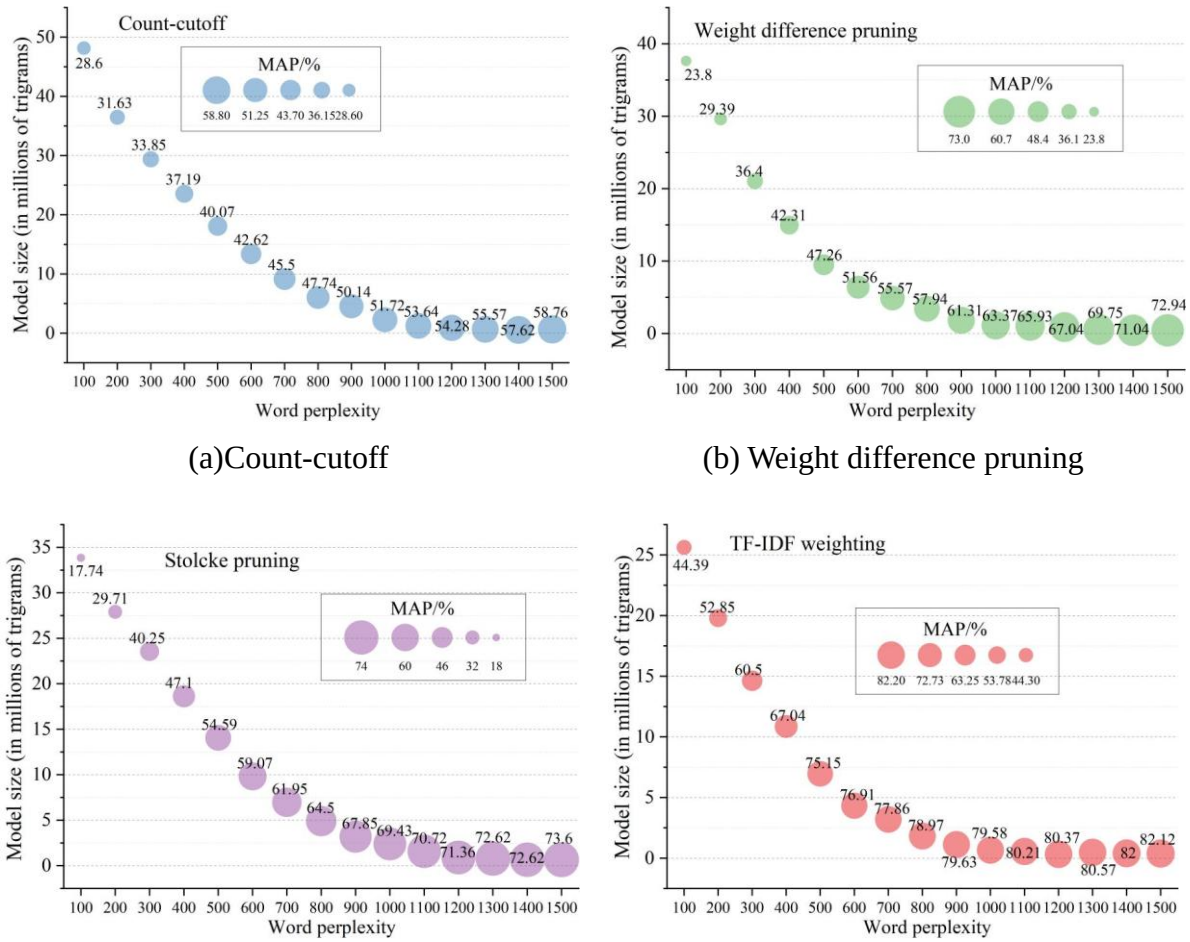
Then, from the perspective of model compression, we examine the impact of this paper's TF-IDF weighting strategy on the final retrieval performance and model size after the introduction of word perplexity as a feature screening criterion. The following strategies are selected as comparison compression techniques, all of which use Bayesian smoothing.

Count cropping: simply remove word strings with less than cutoff occurrences from the model, i.e., assume that the number of occurrences of these strings is 0, and then hand over to the smoothing algorithm for smoothing;

Weight difference trimming: after valuation of all trigram strings any N lowest contributing strings can be selected for deletion;

Stolcke cropping: the entropy increase induced by deleting the word string xyz is employed from an information theoretic point of view.

The mean average precision (MAP) is used as an index to test the model retrieval performance, and the model size and retrieval performance under different strategies are shown in Fig. 4. The four subplots of Fig. 4 are all bubble plots, with the vertical coordinates indicating the model size for million trigrams, the horizontal coordinates indicating different word perplexity, and the bubble size indicating the mean average accuracy percentage (MAP).



(a)Count-cutoff (b) Weight difference pruning

(c) Stolcke pruning (d) TF-IDF weighting

Figure 4: Model size and retrieval performance under different strategies

Figure 4 clearly shows how the TF-IDF weighting-based strategy proposed in this paper helps the model find an excellent balance between retrieval power and model size. The experiments control how much vocabulary can be learned by the model by setting the word perplexity from 100 to 1500. As the word-fixation width increases, the model retrieval performance (MAP) generally tends to increase under each strategy, while the model size decreases. At the same word-fixation width, the TF-IDF weighting strategy achieves the lead in basically all. When the word solid width is 500, the TF-IDF weighted model size is only 6.96 million trigrams, while achieving 75.66% MAP, while the other cropping strategies have model sizes ranging from 14.05-18.07 million trigrams, and the retrieval performance is only 40.07%-54.59%. This indicates that other methods may crop out relevant information while compressing the model size making the model less accurate, whereas TF-IDF weighting achieves efficient information encapsulation by emphasizing discriminative words.

3.2 Lexical analysis of Spanish with additional linguistic rules in statistical modeling

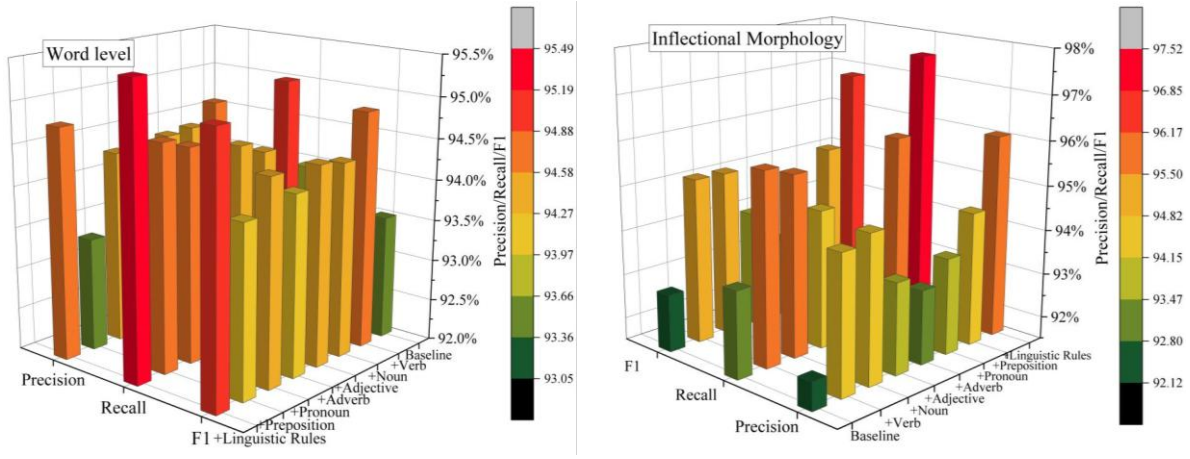
The previous section verified that the Bayesian smoothing-based processing and TF-IDF weighting strategy can help Statistical Language Modeling (SLM-IR) efficiently locate sentences related to preposition usage. In order to deeply analyze the root cause of bias generation, it is necessary to go deeper than the sentence or word level to the internal word

construction rules and grammatical relations. In this regard, this section turns to lexical analysis. Spanish is a language rich in morphological variation, and a word form itself contains key grammatical information such as person, number, tense, and gender. This information directly affects the use of prepositions. Spanish words can be subdivided into stems, hyphenated suffixes, and split suffixes, both of which are linguistic units with a smaller granularity than words. After the lexical analysis system slices words into stems and affixes, not only the lexical and meaning information of the words can be obtained, but also grammatical information such as singular and plural, person and so on can be obtained from the affixes.

In order to achieve a more accurate analysis of the learner's language, lexical analysis systems that incorporate linguistic rules on the basis of statistical models are studied. The experimental evaluation metrics of the Spanish lexical analysis system in this paper include evaluation metrics at the stem/fix unit level. In this paper, correctness, recall and F1 values are used for lexical analysis to evaluate the merit of the system's results.

The corpus used for the experiments is also from the Corpus of Spanish Learners (CORELE), a 100,000-word cut and labeled corpus after manual proofreading. Based on the baseline model, different types of lexical specific rules and integrated linguistic rules are added hierarchically. Lexical-specific rules include the addition of construction and inflection rule bases for verbs, nouns, adjectives, adverbs, pronouns, and prepositions, respectively.

The performance of the study for the statistical linguistics based lexical analysis model on the dataset after adding different lexical and linguistic rules is shown in Figure 5.



(a)Word level (b) Inflectional Morphology

Figure 5: The performance of the lexical analysis model on the dataset

After the baseline model incorporates specific lexical rules and comprehensive linguistic rules, the model improves in all metrics. At the lexical level, the baseline model itself is already robust, with F1 = 93.50. The addition of any class of lexical rules, such as verbs and nouns, gives the model a clearer categorization guideline, and both improve performance by 1-1.5 percentage points. For example, after adding verb rules, F1 rises to 94.87%. It proves that lexical rules can effectively correct the statistical model's fuzzy judgment on the overall attribution of words.

4 Analysis of Chinese learners' Spanish prepositional bias

After verifying the reliability of the model based on Chapter 3, we return to the core research question, the latent cognitive motivations behind Chinese learners' prepositional bias patterns

in Spanish. This chapter provides an in-depth analysis of their prepositional usage in terms of two dimensions: frequency statistics and bias analysis.

Continuing with the Corpus of Learners of Spanish (CORELE) as the research dataset, Spanish text data produced by Japanese, Korean, and English learners as well as native Spanish speakers are obtained.

4.1 Frequency analysis of learners' use of Spanish prepositions

From the perspective of traditional language typology, Chinese is a typical isolating language, Japanese is a typical clinging language, and both Spanish and English are typical flexion languages. These three different types of languages use different grammatical means to express the same grammatical meaning. This typological feature of the native language will influence learners to acquire the target language with a certain bias. This paper focuses on analyzing the use of Spanish prepositions by Chinese learners, and takes the corpus of Japanese learners, Korean learners and English learners as comparative items to explore the reasons why Chinese students develop certain biases in Spanish mediants.

4.1.1 Frequency of use of Spanish prepositions

The frequency of Spanish usage by different learners was first compared. The number of actual occurrences of the prepositions was used as the numerator and the length of the entire corpus was used as the denominator to calculate the share of each item in the set. The 20 Spanish prepositions used with high frequency in the corpus were selected, and the frequency of Spanish prepositions used by students from different native language backgrounds is shown in Table 2.

Table 2: The frequency of Spanish preposition usage among different students

	Spanish	Chinese	Japanese	Korean	English
a	42.83	36.56	31.22	33.83	40.81
de	38.92	40.18	37.61	38.23	35.53
en	21.87	24.52	28.38	27.97	18.46
con	8.94	7.82	5.98	6.55	9.84
para	6.85	9.59	4.83	5.36	10.24
por	6.59	4.77	3.96	4.55	9.61
sin	3.39	2.83	3.51	2.89	2.69
sobre	3.03	3.96	1.88	2.21	2.06
entre	2.73	3.39	2.17	2.44	2.25
desde	2.37	1.75	1.54	1.97	2.81
hasta	2.18	2.81	1.44	1.74	1.67
hacia	1.74	1.29	0.96	1.08	1.47
contra	1.26	0.77	0.85	0.82	1.69
ante	1.02	1.38	0.62	0.79	0.57
según	0.94	0.62	0.37	0.44	0.82
tras	0.90	0.71	0.51	0.60	0.70
durante	2.54	1.98	1.99	1.86	3.83
mediante	0.49	0.28	0.20	0.24	0.42
bajo	1.24	0.86	0.72	0.82	0.75
excepto	0.32	0.21	0.17	0.16	0.37
Total	149.25	146.28	128.91	134.55	146.59
Average	7.46	7.31	6.45	6.73	7.33

The data in Table 2 is the frequency of occurrence per 1,000 words, i.e., number of times/total length of the corpus × 1,000.From the statistics in the table above, it is clear that the use of Spanish prepositions by all the different native learners is lower than that of native Spanish speakers, with the average frequency of use of Spanish prepositions by Chinese, Japanese, Korean, and English learners at 7.31, 6.45, 6.73, and 7.33 words per 1,000 words, respectively, while the Spanish native speakers The average frequency of use for these 20 common prepositions was 7.46 per 1,000 words;

Specifically, the prepositions “a” and “de” are the absolute core prepositions in Spanish, and are used much more frequently than the other prepositions in all groups, with 42.83 and 38.92 per thousand words among native speakers, and 36.56 and 40.18 per thousand words for Chinese learners; Even Chinese learners use it more frequently than native speakers. This might be due to the influence of the Chinese word "de", and they overuse the word "de" which indicates belonging. It is also likely that “en” is used more frequently, influenced by “in”, with a frequency of 24.52 and 21.87 per 1,000 words, respectively.

For English language learners, “para” may be overused due to the English equivalent of “for” with a frequency of 10.24 per thousand words; Meanwhile, due to the influence of “through/during”, “por” and “durante” are also used more frequently, with frequencies of 9.61 and 3.83 (only 6.85 and 2.54 for native speakers). 3.83 (only 6.85 and 2.54 for native speakers). The core preposition “en” is used less frequently than native speakers, at 18.46 per 1,000 words, due to the fact that English tends to use “in/on/at”.

For Japanese and Korean learners with underdeveloped prepositional systems or completely different linguistic backgrounds, there may be a tendency to avoid certain abstract or complex prepositions, so they use them less frequently, such as según, mediante, and rely more on textbook high-frequency words, for example, the frequency of use of the preposition en is between 27-29 per 1,000 words in both groups, compared to 21.87 in the native speakers. In the case of “en”, for example, the frequency of use of this preposition varies between 27-29 per thousand words for both groups of learners, compared to 21.87 for native speakers, probably due to the fact that the spatial marking function of “in” is more direct in their native language, which makes them more inclined to use this preposition with a relatively concrete meaning.

4.1.2 Percentage of prepositional items in total frequency

In order to show more clearly the percentage of prepositional single-item use in the aggregate frequency for different different learners, color mapping was also introduced to visualize the table as shown in Figure 6.

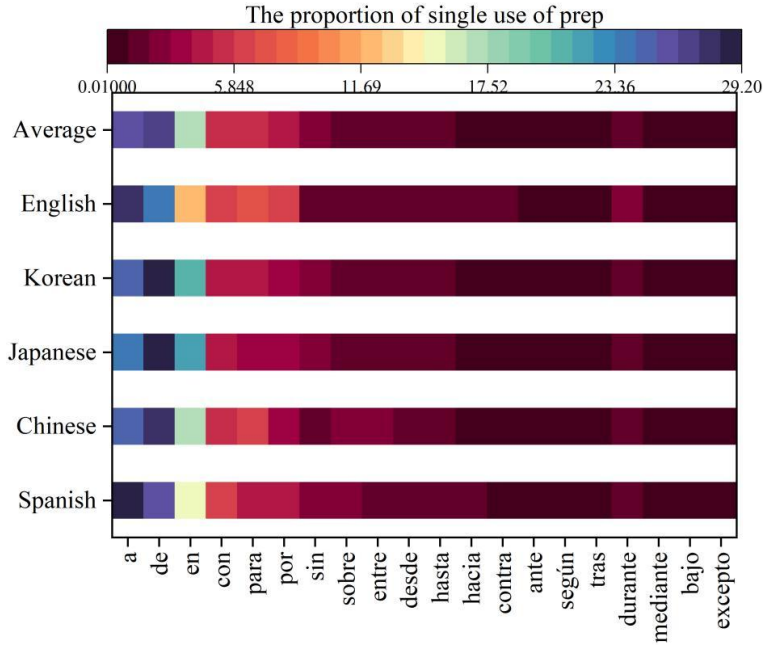


Figure 6: Visualization of the frequency ratio of prepositions usage

Observation of Figure 6 shows that the first three words “a”, "de", and “en” have the darkest color mappings, while the other 17 prepositions are generally clustered in the 0.01-10 red color range. This indicates that the prepositions “a”, "de", and “en” occupy the core structure of Spanish prepositions. At the same time, there are different preferences for different native learners, influenced by their native language transfer.

ELLs use “para” and “por” at 6.99% and 6.56%, respectively, significantly more than all other groups, and even more than native Spanish speakers at 4.59% and 4.42%. This is because it is directly influenced by the multifunctional prepositions such as “for” and “by” in English, leading to a certain degree of overuse. For Chinese learners, their reliance on "de" is the highest among all groups, accounting for 27.47%, while the usage of "a" is relatively low, at 24.99%. It reflects the difference in thinking between the Chinese words "de" and "dui".

The prepositional frequency share once again reveals that prepositional use by different second language learners is always influenced by native language transfer, involving the cognitive habits of the native language in cognitive linguistics.

Further to the in-depth analysis of Chinese learners' Spanish preposition use, the study plotted a Pareto chart of Spanish native speakers and Chinese learners' frequency of use for 20 Spanish prepositions as shown in Figures 7 and 8.

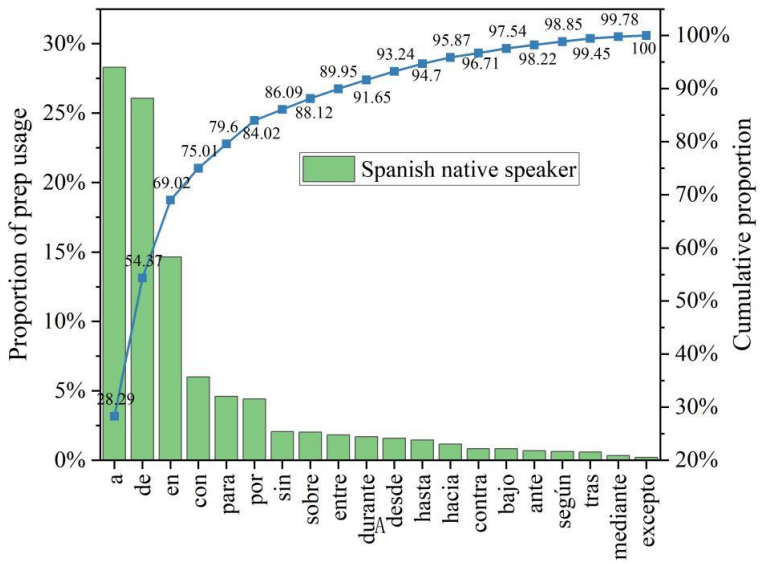


Figure 7: Pareto chart of the frequency of prepositions used by Spanish native speakers

For native speakers, the top three prepositions “a”, “de” and “en” have a cumulative percentage of 69.02%. It means that more than two-thirds of prepositional scenes in real language communication are dominated by these three words. When the scope is extended to the first six words (“a, de, en, con, para, por”), the cumulative share is even 84.02%. From a cognitive point of view, this highly concentrated usage pattern reflects the fact that native speakers develop highly automated mental schemas for these core prepositions. A few words can activate their huge clusters of usage, and it is this ability to master simplicity over complexity that learners need to master.

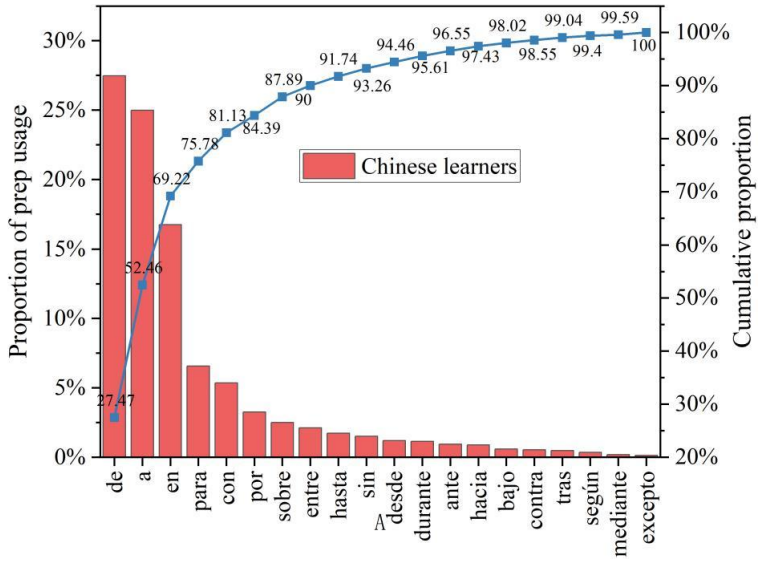


Figure 8: Pareto chart of the frequency of prepositions used by Chinese learners

Comparing the Pareto charts of the Chinese learners with those of the native speakers in Figure 7 reveals that there is a deviation between the two. Although “de” and “a” are also in the top two, the order is reversed, with “de” (27.47%) overtaking “a” (24.99%). This reinforces the observation in Table 3, and may stem from the strong position of the relations of subordination and modification expressed by the Chinese “de” in learners' cognition. More significant differences appear after the third position. Chinese learners use “en” and “para” more than

native speakers. The former were 16.76% and 6.56%, while the latter were 14.65% and 4.59%. The dependence on “en”, as mentioned earlier, may be related to the migration of the spatial marker “in”; The high use of “para”, on the other hand, reflects learners' prioritization of its core denotation of “purpose”, sometimes in place of the more authentic “por” or other constructions.

Observation of the cumulative percentage curve reveals that Chinese learners need to use more prepositions to achieve a cumulative coverage similar to that of native speakers. Native speakers cover 84.02% of the scenes with the first 6 words, while Chinese learners cover 84.39% of the scenes with the first 6 words. Although the values are close, the word order and internal composition are already different. It implies that Chinese learners' expressions are slightly scattered in concentration and have not yet fully internalized the most authentic lexical item choices.

4.2 Statistical analysis of prepositional biases

4.2.1 Analysis of bias rates for different learners

After analyzing the frequency of learners' use of Spanish prepositions, we now turn to the analysis of prepositional biases. Continuing with the CORELE corpus mentioned above, statistics on the occurrence of bias in sentences containing relevant prepositions by different learners are shown in Table 3.

Table 3: The occurrence of errors by different learners of Spanish prepositions

	Chinese	Japanese	Korean	English
a	8.59	10.03	8.30	4.04
de	7.35	7.32	7.48	6.53
en	12.99	19.29	17.66	6.74
con	5.26	7.01	6.09	9.40
para	15.77	8.44	7.66	22.84
por	15.33	9.99	9.17	26.52
sin	4.13	5.69	3.37	3.86
sobre	8.58	6.54	6.09	4.84
entre	3.66	4.06	3.62	4.82
desde	5.06	5.73	4.40	6.56
hasta	5.96	6.12	5.43	4.97
hacia	6.32	7.77	7.57	5.41
contra	5.56	4.74	4.79	7.24
ante	5.88	5.84	5.82	3.09
según	8.51	10.36	10.21	4.17
tras	4.44	4.67	5.08	4.92
durante	7.09	6.08	7.08	12.17
mediante	10.20	11.98	11.47	6.80
bajo	8.06	8.48	7.75	5.75
excepto	9.38	13.34	11.91	4.54
Average	7.91	8.17	7.55	7.76

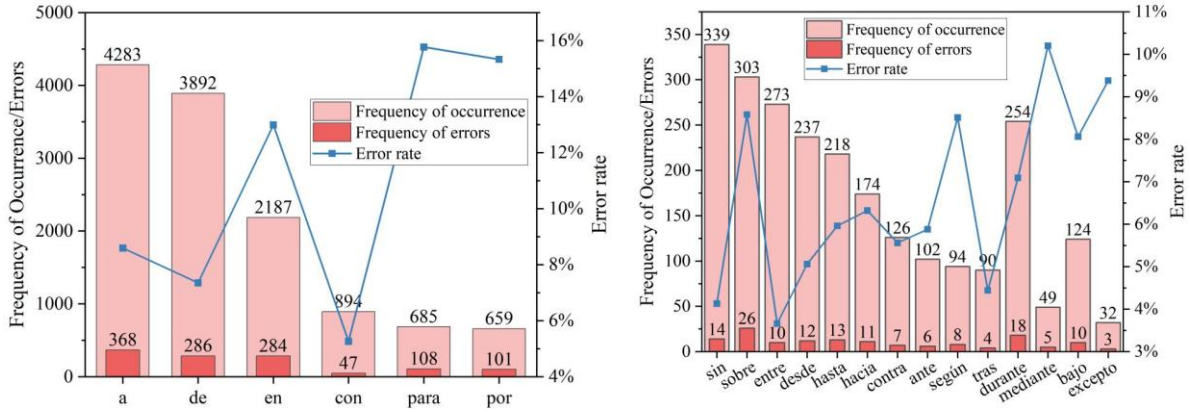
Table 3 reveals the different pain points in the use of Spanish prepositions by learners from different native language backgrounds. Almost all of their usage biases come from negative native language transfer. For English learners, the bias rate is relatively small compared to the

other three types of learners because both English and Spanish are Indo-European systems with many commonalities, but there is still a significant portion of prepositional use bias. It is most prominent in the prepositions “por” and “para”, with a bias rate of 26.52% and 22.84%, far exceeding that of all other groups. It is because in English thinking, learners tend to apply these two multifunctional prepositions with a single schema such as “for” or “by”, which leads to a lot of mis-substitution. The bias rate on “durante” is also significantly higher at 12.17%, which, as mentioned in the analysis above, is likely to be influenced by the usage of English “during”.

For learners from the three East Asian countries, the situation is also very different, with an average bias rate of 7.91%, 8.17% and 7.55% for Spanish prepositions among Chinese, Japanese and Korean learners respectively. Specifically for single prepositions, the three learners together have the highest rate of error for “en”, which is 12.99%, 19.29% and 17.66% respectively. Once again, it is shown that although the native languages are different, the shared locative case marking function of Chinese "in", Japanese "で" and Korean "에" makes learners overly rely on "en" to express various relationships, thereby squeezing the use of other more precise prepositions (such as "a", "de", "sobre"), resulting in generalization bias. They also showed consistently high bias rates for abstract, written, low-frequency prepositions such as “según” and “mediante”, reflecting a general lack of mastery of complex functional prepositions.

4.2.2 Chinese learners' prepositional bias analysis

Now aggregated to Chinese learners, Figure 9 shows the number of times a preposition is used and the number of biases occurring in the written Spanish corpus produced by Chinese learners in the above CORELE. (Since the first 6 core prepositions are not in the same order of magnitude as the others, they are represented in subfigures (a) and (b).)



(a) “a”,“de”, “en”,“con”,“para”“por” (b) “sin”-“excepto”

Figure 9: The frequency and error rate of Spanish prep used by Chinese learners

First of all, high frequency does not equal high bias. The bias rates of “a” and “de”, which are the most frequently used prepositions, are 8.59% and 7.35% respectively, which are in fact at a medium level, indicating that Chinese learners have a good mastery of these two core common Spanish prepositions. The really high bias words are “para” and “por” and “en” with 15.77%, 15.33% and 12.99% bias. The frequency of these three words is not low, with “en” occurring 2,187 times, indicating that learners use them more frequently but do not fully internalize their usage, with more than a 12% probability of making a bias. It is suggested that these complex prepositions need to be specifically attacked. In contrast, Chinese learners' bias rates for such IF words as “entre” and “sin” were smaller, at 3.66% and 4.13% respectively,

with the former entre pointing to a clear space and the latter sin indicating a clearly defined opposition. The former entre (in between) points to a clear space, while the latter sin (for) indicates a clear-cut oppositional relationship. This conceptual clarity reduces the likelihood of confusion with other prepositions, and it is also easy to find a direct correspondence mapping from the learner's native language, i.e., Chinese, which results in positive transfer. In contrast, for por/para/en, which has a high rate of bias, one word covers a variety of abstract relations such as purpose, cause, way, place, etc., which is very prone to misclassification and substitution. These prepositional biases are analyzed next for specific types of bias.

4.2.3 Analysis of Chinese learners' prepositional bias types

The study categorized the prepositional bias types into omission (fewer constituents), addition (more constituents), substitution (improper use, applying the correct substitution down), and misordering (incorrect word order). Table 4 shows the number of occurrences of these 20 preposition-specific bias types.

Table 4: Frequency of error types

	Number of errors	Omissions	Additions	Substitutions	Out-of-order sequences
a	368	92	37	221	18
de	286	80	29	165	12
en	284	71	34	166	13
con	47	8	5	32	2
para	108	11	16	78	3
por	101	10	15	73	3
sin	14	3	2	8	1
sobre	26	4	3	17	2
entre	10	2	1	6	1
desde	12	2	2	7	1
hasta	13	3	1	8	1
hacia	11	2	1	7	1
contra	7	1	1	4	1
ante	6	1	1	3	1
según	8	1	1	5	1
tras	4	1	0	2	1
durante	18	2	3	12	1
mediante	5	1	1	2	1
bajo	10	2	1	6	1
excepto	3	1	0	1	1
Total	1341	299	154	830	58
%	100%	22.30%	11.48%	61.89%	4.33%

Substitution is the most significant type of bias, occurring a total of 830 times, accounting for 61.89% of the total. Among them, the mutual substitution between "por" and "para", as well as the improper use of "en" to replace "a" or "de", constitute this part of the main bodySecondly, omissions account for 22.30%, and this phenomenon is highly concentrated in the three high-frequency prepositions "a", "de", and "en", with these three types of errors accounting for 25% to 27.97% of the total errors. This is often because learners fail to fully internalize certain mandatory syntactic structures in Spanish (such as the requirement for "a" in "ir a + location")

or highly exaggerated usages (such as "de" in the table), and are influenced by places in their native Chinese where prepositions are not needed, thus omits them. In contrast, additions and misordering accounted for a smaller percentage. 11.48% and 4.33% respectively. Addition is even influenced by second language acquisition of English, which is the second language that Chinese learners generally learn from childhood, and tends to influence learners subconsciously, e.g., by the influence of English “for”, adding more prepositions where “para” is not needed. The smallest proportion of “wrong order” is due to the fact that the order of Spanish prepositional phrases is relatively fixed, usually preposition + noun, which is similar to that of Chinese and not easy to make mistakes.

5 Conclusion

Incorporating comprehensive linguistic rules on the basis of statistical modeling, the accuracy of lexical analysis is significantly improved, especially in the fine-grained stemmer-affix level analysis, where the F1 value is increased from 92.79% to 96.78% in the baseline.

Chinese learners' prepositional use is structurally unbalanced compared to native Spanish speakers. Native speakers express the cornerstone of “a” with a frequency of 42.83%, while Chinese learners use “de” most frequently, with 27.47%. This excessive highlighting of "de" and the relative weakening of "a" are the projection of the strong grammatical function of "de" in Chinese in Spanish. Meanwhile, the frequency dependence on “en” (16.76%) and “para” (6.56%) is also higher than that of native speakers, suggesting that learners rely on the spatial schema of ‘in’ and the purpose schema of “in order to” as cognitive mechanisms for understanding Spanish-related categories. This suggests that learners use the Chinese “in” spatial schema and “in order to” purpose schema as the cognitive mechanism for understanding the relevant categories in Spanish.

The high bias rates of 12.99% and 15.77% for “en” and “para” respectively indicate that these prepositions, which are used frequently, have the most confusing conceptual boundaries. Of all the 1,341 cases of bias, substitution bias accounted for 830 cases, or more than 60%. For example, when expressing “in a certain way”, “por” may be wrongly chosen instead of “mediante”.

Research links surface language bias to deep cognitive genesis. Learners' biases are subconscious choices driven by cognitive habits in their native language. This suggests that Spanish language teaching interventions should go deeper into conceptual categories and imagery schema reconstruction in order to more effectively guide learners to build mental representations that fit the cognitive model of the target language.

About the Author

Dexin Kong was born in Xi’an, Shaanxi, P.R. China, in 1996. Currently she is a spanish lecturer at Xi’an Fanyi University. Her main research interests include Spanish language and Spanish literature.

References

[1] Vestergard, T. (2019). Prepositional phrases and prepositional verbs: a study in grammatical function (Vol. 161). Walter de Gruyter GmbH & Co KG.

[2] Tulabut, R. J., Guzman Jr, R. V., Abaring, P. R. M., Armada, A. P., Ilustre, A. H., & Torda,

M. J. T. (2018). Common Errors in Prepositions Committed by Grade 9 Students: Implications for Teaching. TESOL International Journal, 13(3), 113-123.

[3] Fontaine, L. (2017). On prepositions and particles: a case for lexical representation in systemic functional linguistics. Word, 63(2), 115-135.

[4] Otani, N., & Hollmann, W. B. (2025). From within and beyond: a Cognitive Grammar analysis of prepositional complements of prepositions. Cognitive Linguistics, 36(3), 499-526.

[5] Rubba, J. (2011). Grammaticalization as semantic change: a case study of preposition development. In Perspectives on grammaticalization (pp. 81-101). John Benjamins Publishing Company.

[6] Sanjaya, A. A., & Bram, B. (2021). Investigating preposition usage problems of English language education study program students. SAGA: Journal of English Language Teaching and Applied Linguistics, 2(1), 19-34.

[7] Kovbasko, Y. (2021). Grammatical Approaches to Prepositions, Adverbs, Conjunctions, and Particles in Late Modern English. Kalbų Studijos, (38), 99-114.

[8] Hsiung, H. J. (2023). Chinese locative expressions: prepositions and localizers. In Chinese language resources: data collection, linguistic analysis, annotation and language processing (pp. 357-382). Cham: Springer International Publishing.

[9] Cheung, C. C. H. (2016). Chinese: Parts of speech. The Routledge encyclopedia of the Chinese language, 242-294.

[10] Peck, J., & Lin, J. (2019). Semantic constraint on preposition incorporation of postverbal locative PPs in Mandarin Chinese. Language and Linguistics, 20(1), 85-130.

[11] Ursini, F. A. (2013). On the syntax and semantics of Spanish spatial prepositions. Borealis–An International Journal of Hispanic Linguistics, 2(1), 117-166.

[12] Li, J., & Cai, J. (2016). L1 transfer in Chinese learners' use of spatial prepositions in EFL. New Perspectives on Transfer in Second Language Learning, 92, 63.

[13] Perpiñán, S. (2015). L2 grammar and L2 processing in the acquisition of Spanish prepositional relative clauses. Bilingualism: Language and Cognition, 18(4), 577-596.

[14] Santo, A. B. E., & Santo, A. E. (2024). L1 P-Chopping and L2 Null-Preposition: the same output, a different nature. Glossa: a journal of general linguistics, 9(1).

[15] Pamies-Bertrán, A., & Yuan, W. (2020). The spatial conceptualization of time in Spanish and Chinese. Yearbook of Phraseology, 11(1), 107-138.

[16] Concepción Company Company. (2019). Grammatical Words and Spreading of Contexts: Evidence from the Spanish Preposition a. Languages, 4(1), 10.

[17] Torrijos, M. B. (2019). Errors in the use of spanish language: Prepositions. Educación y Futuro Digital, (18), 43-62.

[18] Kissling, E. M. (2024). More evidence that a usage-based, applied cognitive linguistics approach is effective for teaching the Spanish prepositions por and para. Pedagogical Linguistics, 5(1), 1-30.

[19] Lam, Y. (2018). The acquisition of prepositional meanings in L2 Spanish. Canadian Journal of Applied Linguistics, 21(1), 1-22.

[20] Boieblan, M. (2023). Enhancing English spatial prepositions acquisition among Spanish learners of English as L2 through an embodied approach. International Review of Applied Linguistics in Language Teaching, 61(4), 1391-1420.

[21] Eldredge, D. L. (2014). Teaching Spanish, My Way. Xlibris Corporation.

[22] Heydari, P., & Bagheri, M. S. (2012). Error analysis: Sources of L2 learners' errors. Theory and practice in language studies, 2(8), 1583-1589.

[23] Sun, Y., Díaz, L., & Taulé, M. (2019). The development of dynamicity in the acquisition of Spanish by Chinese learners. ITL-International Journal of Applied Linguistics, 170(1), 79-110.

[24] Casañ-Pitarch, R., & Gong, J. (2021). Testing ImmerseMe with Chinese students: acquisition of foreign language forms and vocabulary in Spanish. Language Learning in Higher Education, 11(1), 219-233.

[25] Feng, Y., Iriarte, F., & Valencia, J. (2020). Relationship between learning styles, learning strategies and academic performance of Chinese students who learn Spanish as a foreign language. The Asia-Pacific Education Researcher, 29(5), 431-440.