Quantifying Discourse Support for Omitted Pronouns

Shulin Zhang[1], **Jixing Li**[2], **John Hale**[1]
[1]University of Georgia, US
[2]City University of Hongkong, China
shulin.zhang@uga.edu
jixingli@cityu.edu.hk
jthale@uga.edu

Abstract

Pro-drop is commonly seen in many languages, but its discourse motivations have not been well characterized. Inspired by the topic chain theory in Chinese, this study shows how character-verb usage continuity distinguishes dropped pronouns from overt references to story characters. We model the choice to drop *vs.* not drop as a function of character-verb continuity. The results show that omitted subjects have higher character history-current verb continuity salience than non-omitted subjects. This is consistent with the idea that discourse coherence with a particular topic, such as a story character, indeed facilitates the omission of pronouns in languages and contexts where they are optional.

1 Introduction

Pro-drop is a phenomenon that pronouns can be omitted when they are inferable. It is common across the world's languages, and Mandarin Chinese is one of them (See examples (3) and (4) in Figure 1). Omitted pronouns in these languages, also called zero pronouns, are increasingly important in computational linguistics (*e.g.* Chen et al., 2021; Iida et al., 2006, 2015; Kong et al., 2019). This paper formalizes the notion of Topic Chains, introduced by Tsao (1977) and demonstrates that people omit pronouns when a certain kind of discourse salience is high. We show that this notion of salience is robust across various choices of language models, however, locality (*i.e.* clause recency) seems to be a key requirement.

The proposed formalization leverages the idea that verbs predicated on the same story-character exhibit discourse coherence (Huang, 1984, 1994; Li and Thompson, 1979). Figure 1 shows a literary example where the same character, the narrator, is explicitly referred to once using an explicit pronoun "wo". After that, the pronoun is dropped. The list of predicates (shown in red) applying to the narrator in examples (1) - (3) is ["draw", "lose", "draw"]. When faced with another omitted pronoun in example (4), the fact that the predicate is also "draw" supports the interpretation that the omitted element refers to the narrator. This is because "draw" is similar to the history verbs "draw" and "lose" which were predicated of this same character earlier in the discourse. In this short example, there are other entities such as "grownups" and the "boa constrictor", but their verb histories make them less plausible as candidate referents of the omitted pronoun.

In this paper, we use representations from three neural language models to quantify character-verb usage continuity in a literary discourse, and calculate salience values for each of 32 possible characters at the site of each omitted pronoun. Figure 2 summarizes the analytical steps of this process. Our contributions are as follows: (1) We provide a numerical description of the topic chain continuity. (2) We elaborate on the role of verbs in resolving omitted pronouns. (3) We show that verb similarity and clause range offer reliable clues about the referent of the omitted pronoun.

2 Related Work

Various linguistic theories point to discourse coherence as a factor that enables or encourages *pro*-drop. One of these is Tsao's (1977) notion of Topic Chain. As reviewed in Pu (2019a), a topic chain is a sequence of clauses sharing an identical topic that occurs overtly in one of the clauses. Topic Chains may cross several sentences and even paragraphs (Li, 2004). The multiclausal aspect of Topic Chains supports

Proceedings of the Fifth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2022) , pages 1–12
Gyeongju, Republic of Korea, October 16–17, 2022.

(1) 这 是 我 给 他 后来 画 出来 最好 的 一 幅 画像。
zhe shi wo gei ta houlai hua chulai zuihao de yi fu huaxiang
This is I for he later draw out best DE one drawing
"This is the best portrait I drew for him later on."

(2) [我] 六 岁 时, 大人们 使 我 对 我的 画家 生涯 失去 了 勇气。
wo liu sui shi darenmen shi wo dui wode huajia shengya shiqu le yongqi
[I] Six year old grown-ups make I towards my painter career lose LE courage
"When I was six, grown-ups made me lose courage in my painter career."

(3) [我] 除了 画 过 开着 肚皮 和 闭着 肚皮 的 蟒蛇,
wo chule hua guo kaizhe dupi he bizhe dupi de mangshe
[I] except draw PASS opening belly and closing belly DE boa
"Except that I had drawn boas with opening and closing belly,"

(4) [我] 后来 再 没有 学 过 画。
wo houlai zai meiyou xue guo hua
[I] afterwards again not learn PASS draw
"I had never learned drawing afterwards."

Figure 1: Example of Chinese omitted pronouns in a topic chain. Omitted pronouns, shown here in green with square brackets are not actually spoken. However, their intended reference is unambiguous for native speakers. Predicates are shown in red, and the overtly expressed entities are shown in blue. Unlike in Romance languages, there is no morphological change on verbs to mark the gender or number of omitted elements in Chinese.

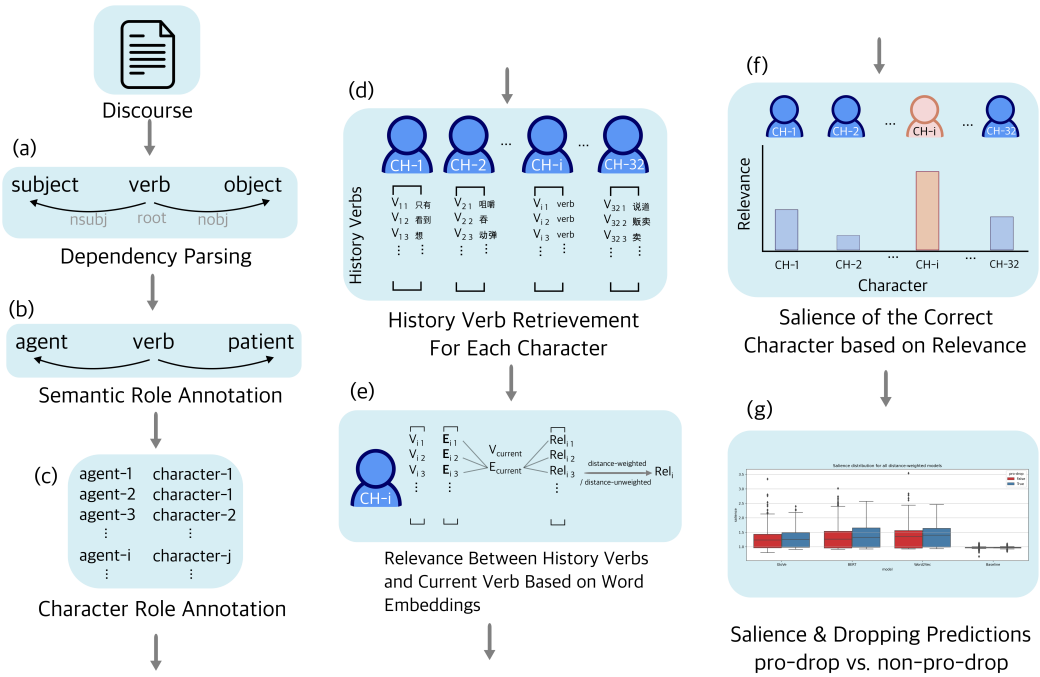

Figure 2: Analysis steps adopted in this study: (a) Grammatical subjects and objects of each main verb are identified via dependency parsing on the whole story discourse of The Little Prince (See a sentence example from Table 1, columns "S", "V", "O"); (b) Semantic role annotation: for all the subjects and objects, annotate their semantic roles as AGENT or PATIENT (See Table 1 column "V-agent" and "V-patient"); (c) Character role annotation: assign story character roles to the entities, see character occurrences in Table A1, and Table 1 column "character"; (d) History verb retrieval for each story character: for each story character, tabulate the verbs that are its main verbs being used in the discourse (See example Table A3); (e) Relevance between history verbs and a current verb: for each current verb, calculate its relevance to the history verbs, and sum with or without their distance weight (See Table 2 and A5); (f) Salience of the correct character: for each verb, calculate how "salient" the correct character is compared to all other characters (See example Table A6); (g) Group test between *pro*-drop verbs vs. non-*pro*-drop verbs, and apply logistic regression to test predictability of character salience on dropping behavior (See group results in Table 3 and Figure 3).

long-distance coreference (Sun, 2019). Taking a dynamic perspective, Pu (2019b) suggests that a topic chain "encodes a referent that is cognitively most accessible at the moment of discourse production, as enhanced by maximum discourse coherence of topic continuity and thematic coherence".

We conceptualize accessibility in Pu's sense as the relative salience of a story character that participates in a chain of predications. Instead of focusing on named entities, we form the chain based on the verbs in the preceding discourse.

3 Method

3.1 Discourse Material

The discourse material used in this study is a Chinese translation (xiaowangzi.org, 2021) of Saint-Exupéry's *The Little Prince*. It contains 2802 clauses and 16010 words, and the word tokenization was manually checked by native Chinese speakers.

3.2 Dependency Structure Retrieval and Semantic Role Annotation

We manually annotate the semantic roles Agent and Patient for each verb using dependency analyses provided by Stanza (Qi et al., 2020) and part of speech tags provided by spaCy. For most cases in the discourse, subjects are acting as agents whereas objects are acting as patients, but there are 494 exceptions (*i.e.* 218 Agents are acting as Objects, and 276 Patients are acting as Subjects) such as passives, the -BA('把') construction, the relative clause -DE('的') construction *etc.* that call for our manual annotation (See Chapter 28 and 32 in The Oxford Handbook of Chinese Linguistics (Wang and Sun, 2015) regarding these constructions).

The textual antecedents of each agent and patient are separately annotated manually. As shown in Table 1, the sentence meaning "These boas swallow their prey without chewing" has the following annotations: verbs annotated in column *V*; verbs' agents and patients annotated correspondingly in column *V-agent* and *V-patient*; pronouns or named entities' character roles are annotated in column *character*. As described below in Section 3.4, information about characters in particular semantic roles can be used to form a dynamic usage table, reifying Pu's view of Topic Chains.

ID	word	S	V	O	V-agent	V-patient	character
56	这些 (these)						
57	蟒蛇 (boa)	True					ch2_boa
58	把 (BA)						
59	它们 (them)						
60	的 (DE)						
61	猎获物 (prey)			True			
62	不 (not)						
63	加 (with)						
64	咀嚼 (chew)		True		57 (boa)	61 (prey)	
65	地 (DI)						
66	囫囵 (roughly)						
67	吞 (swallow)		True		57 (boa)	61 (prey)	
68	下 (down)						

Table 1: Dependency structure and semantic role annotation table. An annotation example for the sentence "These boas swallow their prey without chewing." The verbs "chew" and "swallow" are located as verbs in the column *V*. Token indices for each verb's Agent and/or Patient are annotated in the columns *V-agent* and *V-patient* respectively, and the character roles they are referring to are annotated in the column *character*.

3.3 *Pro*-drop Annotation

Omitted subjects and objects are manually resolved using numerical indices from 1 to 32. As shown in Appendix Table A2, 422 Agents and 16 Patients are found omitted in the discourse, and in the following analyses, we focus on just story characters in the Agent semantic role.

3.4 Dynamic Character-Verb Usage Table

Based on the dependency annotation table, the verbs used for each character are extracted and entered in a second table, the Character-Verb Usage Table (See example in Appendix Table A3). This table includes the following features: (1) *verb*, the original text of the verb; (2) *verb_id*, the index of the verb in the whole discourse; (3) *agent/patient_character*, the verb's agent or patient story character; (4) *pro_drop*, whether the verb has *pro*-drop; (5) *ch[1-32]_prev_verbs*, for characters 1 through 32, their corresponding previous verbs and indexes are stored as lists.

The dynamic character-verb usage table includes the previous verbs for each story character until a "current verb", and this indicates the verb usage history of each character. By transforming these verb usage histories into numerical vectors, it is possible to use a sim-

ple notion of similarity to formalize discourse coherence.

3.5 History-verb and Current-verb Relevance

The idea behind comparing the history verbs and the current verb for each story character is to calculate a numerical similarity level between the current verb and preceding verbs that are part of one or another Topic Chain. Inspired by Sperber and Wilson (1986), we define a quantity called Relevance, a time-weighted function of vector similarity with preceding predicates. The Relevance evaluation process adopt three types of word embeddings (See Section 3.5.1 for details), and steps for the evaluation are introduced in Section 3.5.2.

3.5.1 Word Embeddings Methods

Word embeddings allow each word to be mapped to a single point in a vector space. Under the Distributional Hypothesis (see *e.g.* Lenci, 2018), words with similar meanings should be closer in vector space (for a textbook introduction, see Pilehvar and Camacho-Collados, 2020). We use this idea to calculate the similarity between the main verb of an omitted pronoun and the verb chains of story characters that might serve as that omitted pronoun's referent.

We use three types of word embeddings: GloVe, BERT, and Word2Vec. The GloVe model (Pennington et al., 2014) learns word embedding from the term co-occurrence matrix by minimizing the reconstruction error. GloVe has a large context window, which allows it to capture longer-term dependency features. The BERT model (Kenton and Toutanova, 2019) consists of multi-layer bidirectional transformer encoders. BERT is trained on two unsupervised tasks: predict masked tokens, and predict the next sentence, and the BERT embeddings reflect contextual corpus features. Word2Vec is a prediction-based model (Mikolov et al., 2013a,b), and the word embeddings used in this study (Li et al., 2018) were trained on a Skip-Gram with Negative Sampling (SGNS) model. All word embeddings we used were trained on large Chinese corpora, and contain contextual word knowledge that carries semantic, syntactic, and pragmatic features. Among these three word embedding models, BERT can provide contextualized features of the language compared to the others due to the tasks and processes it has been trained on.

In this study, BERT and GloVe models are applied with spaCy[1], and Word2Vec model is applied with pretrained Chinese Word Vectors[2] (Li et al., 2018). A baseline model with 300-dimension random value vectors is adopted to calculate the baseline relevance as compared to the other word embedding models.

The GloVe word embeddings are obtained from the *zh_core_web_lg* model in spaCy. The GloVe model (Pennington et al., 2014) relies on word co-occurrence in the training corpus, and considers the ratios of word-word co-occurrence probabilities to encode semantic information. The model in spaCy was trained on OntoNotes 5, CoreNLP Universal Dependencies Converter, and Explosion fastText Vectors. It has 500,000 unique vectors with a dimension size of 300. We obtained the word vectors by searching up the Chinese word in the word dictionary.

The BERT word embeddings are retrieved from the *zh_core_web_trf* model in spaCy. This transformer model was trained on OntoNotes 5, CoreNLP Universal Dependencies Converter, and bert-base-chinese. The word embedding vectors were obtained by grouping every 50 words in the discourse, and the model inputs were the 50 words combined as a string (with space between the words). The dimension of the BERT word embedding is 768. If there were more than 1 character in a word, their vectors' mean value was used as the word embedding for the whole word. For example, the word "只有" 's embedding was calculated by averaging its subwords' embedding vectors of "只" and "有".

The Word2Vec word embeddings were pretrained on Word2Vec model with a large Chinese corpus containing data from Baidu Netdisk (22.6G), and the vector dimension is 300 (Li et al., 2018).

Baseline Word Vectors were 300-dimension

[1] https://spacy.io/models/zh

[2] https://github.com/Embedding/Chinese-Word-Vectors

vectors generated randomly in the range -1 to 1. The same analysis steps are applied to this model as a baseline.

3.5.2 Relevance evaluation

The relevance between history verbs and current verbs is calculated based on their word embedding similarities (see Section 3.5.1 for details). At the same time, a weight decay function is applied to the influence of each history verb based on its distance to the current verb, and the function used here is a vanilla value decreasing function (see Equation 1), in which ω refers to the weight applying on the similarity, d refers to the clause distance between the verbs being compared, and j, k are the clause numbers the verbs are in:

$$\begin{aligned} \omega(j,k) &= 1/(d+1) \\ d &= |j-k| \end{aligned} \tag{1}$$

In this study, the "word embedding similarity" method is realized by calculating the Cosine Similarity between two word embedding vectors. As shown in Equation 2, v_{prev} refers to a word embedding vector of a previous verb, and v_{curr} refers to the one for the current verb:

$$R(v_{prev}, v_{curr}) = \frac{v_{prev} \cdot v_{curr}}{||v_{prev}||||v_{curr}||} \tag{2}$$

Therefore, the clause-distance-weighted similarity between history verbs and the current verb is shown as Equation 3, in which n refers to the number of verbs in the history verb list for a character, and *cl_prev_i* and *cl_curr* refer to the clause numbers that the previous verb and the current verb are in correspondingly.

$$\begin{aligned} &R_{weighted}([v_{prev\_1}, ..., v_{prev\_n}], v_{curr}) = \\ &\sum_{i=1}^{n} \omega(cl\_prev\_i, cl\_curr) * R(v_{prev\_i}, v_{curr}) \end{aligned} \tag{3}$$

Via Equation 3, for a current verb, each story character has a corresponding relevance value: if the value is higher, the distance-weighted word embedding similarity between history verbs and current verb is higher; and vice versa.

Appendix Table A3 shows an example of a verb and the history verbs for characters 1 through 32. The GloVe, BERT, Word2Vec, and Baseline embeddings are used to calculate the average relevance of the history verbs to each current verb for each story character.

Regressors obtained from relevance evaluation introduced in this section are shown in Table 2. The average similarity is calculated following Equation 2 and 3. Both distance-weighted and distance-unweighted relevance are explored to see whether clause distance would play a role.

Regressor Number	Regressor Name	Regressor Meaning
1	verb	the verb in the discourse acting as a main verb of a clause
2	verb-id	the word order id of this verb in the original discourse
3	agent-character	the story character referred by the agent of the verb
4	pro-drop	whether this agent is dropped in the discourse
5 - 36	ch{1-32}-prev-verbs	the previous verbs used by each story character till the current verb
37 - 68	rel-glove-ch{1-32}	relevance obtained by GloVe word embeddings
69 - 100	rel-bert-ch{1-32}	relevance obtained by BERT word embeddings
101 - 132	rel-word2vec-ch{1-32}	relevance obtained by Word2Vec word embeddings
133 - 164	rel-baseline-ch{1-32}	relevance obtained by Baseline word vectors

Table 2: Regressors obtained after the relevance calculation

As shown in Appendix Table A5, the relevance calculation results of the last verb are presented as an example.

3.6 Character Salience

With the relevance between history-current verbs computed as described in the previous section, we have a similarity value for each story character to the current verb. This character salience value refers to whether a story character stands out compared to other candidate characters. The salience level function is described in Equation 4. In Equation 4, k refers to character_k, and the relevance values were calculated based on its history-current verbs by Equation 3.

$$S(k) = \frac{\sum_{i=1}^{n} \left(\frac{R_{weighted}(k)+1}{R_{weighted}(i)+1} \right)}{n+1} \tag{4}$$

3.6.1 Ranged Character Salience

Instead of taking all 32 story characters as candidates for the salience value calculation, the

		Correct character salience pro-drop >non-pro-drop (n = 422)							
Candidates' Range		**range = all**		**range <10 clause**		**range <20 clause**		**range <30 clause**	
		t-value	*p-value*	*t-value*	*p-value*	*t-value*	*p-value*	*t-value*	*p-value*
Distance-Weighted	**GloVe**	49090.319	0.063	51137.593	0.012*	52598.233	0.003**	52121.241	0.004**
	BERT	50555.45	0.023*	45310.076	0.029*	52105.854	0.005**	51582.819	0.008**
	Word2Vec	50358.954	0.025*	51268.800	0.011*	52747.81	0.002**	52246.569	0.004**
	Baseline	44656.318	0.496	44737.336	0.483	49199.853	0.060	47875.291	0.134
Distance-Unweighted	**GloVe**	39345.494	0.959	44384.169	0.531	43818.383	0.606	43837.85	0.604
	BERT	42867.41	0.724	45310.076	0.411	45187.343	0.425	45220.75	0.421
	Word2Vec	40865.782	0.898	45236.126	0.420	44672.755	0.494	44630.117	0.498
	Baseline	43149.674	0.690	45940.625	0.330	46398.831	0.275	45552.563	0.377

Table 3: Single-sided nonparametric two-sample Wilcoxon test between *pro*-drop and non-*pro*-drop salience values among three word embedding models and the baseline model: With candidates included as all candidates, candidates within 10 clauses, 20 clauses, and 30 clauses.

		Logistic Regression Model Pro-drop Prediction Accuracy			
Candidates' Range		**range = all**	**range <10 clause**	**range <20 clause**	**range <30 clause**
Distance-Weighted	**GloVe**	**0.518**	**0.535**	**0.527**	**0.539**
	BERT	**0.538**	**0.532**	**0.536**	**0.546**
	Word2Vec	**0.534**	**0.535**	**0.537**	**0.552**
	Baseline	0.497	0.489	0.495	0.498
Distance-Unweighted	**GloVe**	**0.524**	0.487	0.490	0.485
	BERT	0.493	0.488	0.492	0.482
	Word2Vec	**0.514**	0.485	0.482	0.473
	Baseline	0.485	0.485	0.485	0.485

Table 4: *Pro*-drop prediction accuracy results of the Logistic Regression model from three word embedding models and one baseline model: salience value calculated based on all previous clauses and ranged clauses.

ranged candidates' salience compares the correct character's accumulated relevance value to the ones within a certain number of clauses. We consider candidates within 10, 20, and 30 clauses for this ranged salience.

3.7 *Pro*-drop Prediction

With the correct story character's salience level for each verb in the annotated discourse, we apply a logistic regression model to predict *pro*-drop based on the salience level. The sample sizes are chosen by the size of the smaller group (*i.e. pro*-drop), and the chosen processes are repeated 100 times to obtain the average accuracy level.

4 Results

In this study, the following analyses are applied to The Little Prince discourse to explore the effect of verb continuity on the *pro*-drop phenomenon: (1) relevance between history and current verbs for all story characters (with three types of word embeddings applied); (2) character salience of the correct character, and (3) correct character salience group comparison, and its predictability on *pro*-drop in the discourse. The following sections describe the results of (2) and (3), and (1) is an intermediate step introduced in Section 3.5.

4.1 Character Salience: *Pro*-drop vs. Non-*pro*-drop

The correct story character's salience compared to all other characters was calculated following Equation 4. For each verb, we calculated a salience value for the correct story character. See Appendix Table A6 for an example of the salience values of the last verb.

The distributions for the salience value obtained from three word embedding models and one baseline model are shown in Figure 3.

Single-sided nonparametric two-sample Wilcoxon Tests are carried out between *pro*-drop and non-*pro*-drop character salience of

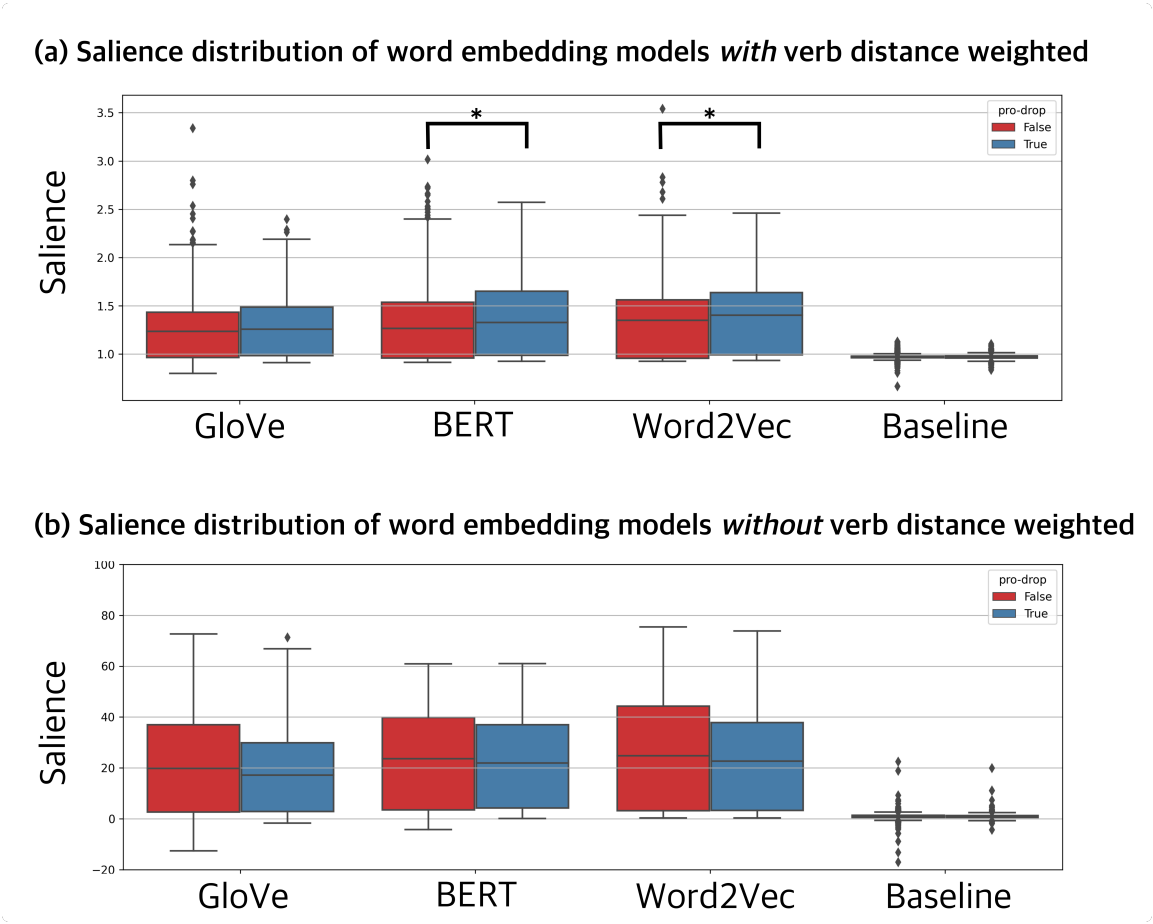

Figure 3: Salience distributions from the word embedding models: GloVe, BERT, Word2Vec, and Baseline. (a) Salience distribution based on distance-weighted models; (b) Salience distribution based on distance-unweighted models. The blue boxplots are *pro*-drop salience cases, and the red ones are non-*pro*-drop. The BERT and Word2Vec models show significant *pro*-drop > non-*pro*-drop effect, and GloVe model shows marginally significant result (See detailed Wilcoxon tests results in Table 3).

all three word embedding models and the baseline model. 422 cases are randomly selected from non-*pro*-drop salience values to match the size of the *pro*-drop ones, and this process is repeated 1000 times to gain the average statistic values. The test results are shown in Table 3. For distance-weighted models, BERT and Word2Vec show significant results ($p < 0.05$), and GloVe show marginally significant result ($p = 0.063$). For distance-unweighted models, none of them show significant results. The Wilcoxon test results based on ranged salience are shown in Table 3 in columns "range < 10/20/30 clauses" as compared to the non-ranged results in "range = all". As shown in the table, the effects of "*pro*-drop > non-*pro*-drop" on correct character salience tend to be larger when the salience is calculated based on ranged clauses. The Baseline model shows null effects on both distance-weighted and distance-unweighted models for all the ranged cases. As shown in Figure 3, the boxplots are consistent with the Wilcoxon tests.

4.2 Logistic Regression: Predict *Pro*-drop with Character Salience

With the salience values described in the previous section, the logistic regression model is applied to examine the effect of salience on *pro*-drop. 75% of the data are used as the training set, and 25% of the data are used as the testing set. See the prediction accuracy results in Table 4 based on salience values obtained from all-ranged and clause-ranged clauses. As shown in Table 4, except for the baseline model, all the distance-weighted language models' results show above chance (> 50%) accuracy.

As for distance-unweighted language models, only GloVe and Word2Vec show above chance results on all-ranged predictions. Similar to the "range-effect" shown in the previous section, it can be seen from the prediction results as well that ranged clauses' prediction accuracies tend to be slightly higher than non-ranged results.

5 Discussion

The main findings of this study are: (1) Compared with overtly expressed subjects, omitted subjects have higher verb-usage continuity. In this respect, they stand out among other story characters; (2) Clause distance plays a role in contextual information strength: With clause distance weighted, the *pro*-drop $>$ non-*pro*-drop salience effects are statistically significant; (3) Constraining the range of candidates by clause recency appears to strengthen these effects.

These results validate Topic Chain theory (Tsao, 1977) by showing how verbs contribute to the discourse coherence that omitted pronouns depend on. The "ranged" recency effect indicates that local contextual coherence might play a more important role than whole-discourse-level coherence. This recency effect may also explain the better performance obtained by BERT and Word2Vec compared to GloVe, since GloVe word embeddings are obtained from discourse-level word co-occurrence statistical features, and BERT and Word2Vec are trained on comparably smaller scale contextual information.

It shall be noted that verb-usage continuity is not the only factor that conditions *pro*-drop. Other factors, including nonverbal lexical information and syntactic patterns *e.g.* with conjunctions, also support discourse coherence (Halliday and Hasan, 1976). In this light, it is remarkable that one factor on its own, verb-usage continuity, yields above-chance accuracy in predicting *pro*-drop.

6 Conclusion

This study quantifies character-verb usage continuity as an aspect of discourse that helps comprehenders resolve omitted pronouns. Omitted pronouns tend to show higher verb usage consistency compared to pronounced entities, and this effect is strengthened by clause recency.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1903783.

References

Shisong Chen, Binbin Gu, Jianfeng Qu, Zhixu Li, An Liu, Lei Zhao, and Zhigang Chen. 2021. Tackling zero pronoun resolution and non-zero coreference resolution jointly. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 518–527.

Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. 1976. Cohesion in English. *English Language Series, Longman, London*.

C-T James Huang. 1984. On the distribution and reference of empty pronouns. *Linguistic inquiry*, pages 531–574.

Yan Huang. 1994. *The syntax and pragmatics of anaphora: A study with special reference to Chinese*. Cambridge University Press.

Ryu Iida, Kentaro Inui, and Yuji Matsumoto. 2006. Exploiting syntactic patterns as clues in zero-anaphora resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 625–632.

Ryu Iida, Kentaro Torisawa, Chikara Hashimoto, Jong-Hoon Oh, and Julien Kloetzer. 2015. Intra-sentential zero anaphora resolution using subject sharing recognition. In *EMNLP*, pages 2179–2189.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Fang Kong, Min Zhang, and Guodong Zhou. 2019. Chinese zero pronoun resolution: A chain-to-chain approach. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(1):1–21.

Alessandro Lenci. 2018. Distributional models of word meaning. *Annual review of Linguistics*, 4:151–171.

Charles N Li and Sandra A Thompson. 1979. Third-person pronouns and zero-anaphora in Chinese discourse. In *Discourse and syntax*, pages 311–335. Brill.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on Chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 138–143, Melbourne, Australia. Association for Computational Linguistics.

Wendan Li. 2004. Topic chains in Chinese discourse. *Discourse Processes*, 37(1):25–45.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2020. Embeddings in natural language processing: theory and advances in vector representations of meaning. *Synthesis Lectures on Human Language Technologies*, 13(4):1–175.

Ming-Ming Pu. 2019a. *Zero anaphora and topic chain in Chinese Discourse*. Routledge.

Ming-Ming Pu. 2019b. Zero anaphora and topic chain in Chinese discourse. In *The Routledge Handbook of Chinese Discourse Analysis*, pages 188–200. Routledge.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and cognition*, volume 142. Citeseer.

Kun Sun. 2019. The integration functions of topic chains in Chinese discourse. *Acta Linguistica Asiatica*, 9(1):29–57.

Fengfu Tsao. 1977. *A functional study of topic in Chinese: The first step towards discourse analysis*. Ph.D. thesis, USC, Los Angeles, California.

William SY Wang and Chaofen Sun. 2015. *The Oxford handbook of Chinese linguistics*. Oxford University Press.

xiaowangzi.org. 2021. 小王子网站. http://www.xiaowangzi.org/. Accessed: 2021-04-03.

A Appendix

Story character	Annotation Label	Number of Occurrence
the little prince	ch4_prince	676
the story teller	ch1_storyteller	356
the rose	ch12_rose	166
the king	ch18_king	71
the fox	ch28_fox	67
the planet	ch11_planet	62
the lamplighter	ch23_lighter	54
the sheep	ch5_sheep	48
the geologist	ch24_geologist	41
the grownups	ch3_grownups	39
the snake	ch26_snake	39
the businessman	ch22_shiyejia	37
readers	ch8_audience	30
the volcano	ch17_volcano	22
the baobab	ch9_tree	20
the drunk man	ch21_drunk	18
the conceited man	ch20_xurong	16
the travelers	ch31_traveler	15
the seed	ch10_grass	13
the explorer	ch25_explorer	13
the red-faced man	ch14_redface	11
the boa	ch2_boa	10
the switch man	ch29_switcher	10
the astronomer	ch6_universescholar	7
the echo	ch27_echo	5
the tiger	ch15_tiger	5
the drafts	ch16_wind	4
the train	ch30_train	4
the merchant	ch32_merchant	4
the children	ch13_kids	3
the general	ch19_general	3
the ruler	ch7_ruler	1

Table A1: The number of occurrence of each character in the annotated discourse

	Agent	Patient
***pro*-drop**	422	16
non-*pro*-drop	2032	1329
total number	2454	1345

Table A2: Distribution of annotated Agents and Patients in the whole discourse.

verb	回来
verb_id	16008
agent_character	ch4
pro_drop	False
ch1_prev_verbs	[只有, 看到, 想, 用, 画, 画, 让, 画,...]
ch2_prev_verbs	[咀嚼, 吞, 动弹, 消化, 消化, 开, 闭, 闭,...]
ch3_prev_verbs	[理解, 看, 懂, 需要, 解释, 劝, 靠, 弄,...]
ch4_prev_verbs	[朝, 望, 出现, 给, 像, 没有, 像, 干,...]
ch5_prev_verbs	[病, 需要, 像, 睡, 去, 用, 跑, 跑,...]
...	...
ch30_prev_verbs	[运载, 发, 往, 朝着, 开, 过]
ch31_prev_verbs	[寻找, 回来, 满意, 住, 追随, 追随, 睡觉, 打哈欠,...]
ch32_prev_verbs	[说道, 贩卖, 卖, 说]

Table A3: Example of Verb-Character table. (See a translation of this table in Table A4)

verb	come back
verb_id	16008
agent_character	ch4
pro_drop	False
ch1_prev_verbs	[have, see, want, use, draw, draw, let, draw,...]
ch2_prev_verbs	[chew, swallow, move, digest, digest, open, close, close,...]
ch3_prev_verbs	[understand, see, understand, need, explain, advise, lean, play,...]
ch4_prev_verbs	[turn, watch, show up, give, alike, (not) have, alike, do,...]
ch5_prev_verbs	[sick, need, alike, sleep, go, use, run, run,...]
...	...
ch30_prev_verbs	[carry, send, go, turn, drive, pass]
ch31_prev_verbs	[look up, come back, satisfy, live, follow, follow, sleep, yawn,...]
ch32_prev_verbs	[speak, sell, sell, say]

Table A4: Translation of Table A3: Example of Verb-Character table.

Relevance Regressor	(Non-weighted relevance, Weighted relevance)
rel_glove_ch1	(81.89066125531684, 0.32419914580071807)
rel_glove_ch2	(1.8756812506219913, 0.001503683756709864)
...	...
rel_glove_ch32	(0.8230171383397842, 0.001262691669193839)
rel_bert_ch1	(176.59183087820725, 0.6119750732174682)
rel_bert_ch2	(4.919826668243348, 0.0027848581443943223)
...	...
rel_bert_ch32	(0.867459723760406, 0.001329274033713714)
rel_word2vec_ch1	(134.572604613474, 0.4595537826115222)
rel_word2vec_ch2	(2.8936049625643223, 0.0020496541891822087)
...	...
rel_word2vec_ch32	(0.9999583161919829, 0.0015334960473239322)
rel_baseline_ch1	(-0.771830408650495, 0.008005141647819333)
rel_baseline_ch2	(-0.008373434318707955, 5.9110606393949324e-05)
...	...
rel_baseline_ch32	(0.08827132539725344, 0.00013526127447238275)

Table A5: Example of relevance results for the last verb

Regressor	Example value
verb	回来 (come back)
correct character	ch4
pro-drop	False
salience-glove-unweighted	45.761057
salience-bert-unweighted	57.886974
salience-word2vec-unweighted	56.125342
salience-baseline-unweighted	1.087911
salience-glove-weighted	1.206085
salience-bert-weighted	1.522071
salience-word2vec-weighted	1.427663
salience-baseline-weighted	0.979743

Table A6: Example of salience results for the last verb from three language models and one baseline model with distance-weighted/-unweighted