Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-024-03317-6 OPEN

Do translation universals exist at the syntactic-semantic level? A study using semantic role labeling and textual entailment analysis of English-Chinese translations

Letao Wang[1,2] & Yue Jiang[1]✉

Albeit extensive studies of translation universals at lexical and grammatical levels, there has been scant research at the syntactic-semantic level. To bridge this gap, this study employs semantic role labeling and textual entailment analysis to compare Chinese translations with English source texts and non-translated Chinese original texts. The research has found substantial evidence for translation universals like explicitation, simplification, and levelling out at the syntactic-semantic level, which is illustrated by significant differences between syntactic-semantic features of Chinese translations and those of English source texts and Chinese original texts. This suggests a distinct syntactic-semantic uniqueness of Chinese translations, wherein the overall features exhibit an "eclectic" characteristic, showcasing contrasting outcomes such as explicitation identified as S-universal and implicitation deemed T-universal. This could be attributed to the gravitational pull from the two language systems. In the inspection of specific semantic roles, features of agents and discourse markers are found to be evidence for both S-explicitation and T-explicitation, potentially reflecting the role of socio-cultural factors in shaping the uniqueness of syntactic-semantic features of Chinese translations. These findings further underscore the complexity inherent in translation, highlighting its function as a dynamic balance system.

[1] School of Foreign Studies, Xi'an Jiaotong University, Xi'an 710049 Shaanxi, China. [2] School of Foreign Studies, Chang'an University, Xi'an 710064 Shaanxi, China. ✉email: jiang8@xjtu.edu.cn

Introduction

The concept of "the third language" was initially put forward by Duff (1981) to indicate that translational language can be distinguished from both the source language and the target language based on some of its intrinsic linguistic features. Frawley (2000) also introduced a similar concept known as "the third code" to emphasize the uniqueness of translational language generated from the process of rendering coded elements into other codes. Baker (1993) then formulated the hypothesis of "translation universals" based on empirical studies of corpora, which also suggests that translation behaviour gives rise to certain universal linguistic features that distinguish the translated texts from both the source texts and original texts in the target language. The question of whether translational language should be regarded as a distinctive language variant has since sparked considerable debate in the field of translation studies. While numerous studies have been conducted to test the translation universal hypothesis and its related sub-hypotheses, most of them have only focused on the lexical and grammatical features in spite that some of the translation universals, such as explicitation and simplification, may be more noteworthy at the semantic and informational level. Given the necessity to involve semantic features for a more systematic study of translation universals, the current study aims to delve into translation universals in English-Chinese translation by employing methods based on semantic role labeling and textual entailment analysis, integrating features at both the syntactic and semantic levels to gain a more comprehensive and in-depth understanding of the translation universal hypothesis.

Literature review

Translation universal hypothesis. Since the translation universal hypothesis was introduced (Baker, 1993), it has been a subject of constant debate and refinement among researchers in the field. On the one hand, some proposed that translation universals can be further divided into T-universals and S-universals (Chesterman, 2004). T-universals are concerned with the intra-linguistic comparison between translated texts and non-translated original texts in the target language while S-universals are concerned with the interlinguistic comparison between source texts and translated texts. On the other hand, some proposed that the hypothesis consists of many sub-hypotheses like simplification (Laviosa, 1998a; Malmkjær, 1997), explicitation (Olohan, 2003; Olohan & Baker, 2000; Øverås, 1998), normalization (Kenny, 2014, 2017), levelling out (Laviosa, 1998b), and the unique item hypothesis (Eskola, 2004; Tirkkonen-Condit, 2004), to name a few. Among these, explicitation stands out to be the most semantically salient hypothesis. It was first formulated by Blum-Kulka (1986) to suggest that translated texts have a higher level of cohesive explicitness. Baker (1996) broadened its definition into the "translator's tendency to explicate information that is implicit in the source text", emphasizing that explicitation in translated texts is not limited to cohesion, but can also be observed at the informational level. Such being the case, measurement of explicitation merely at the syntactic level is not enough, and an investigation of it at the syntactic-semantic level is necessary. Moreover, translation universals like simplification and levelling out reflect the unique characteristics of translational language at the lexical and syntactic level, but they are also likely to cause subtle semantic deviation as well as distortion of the informational structure, which may also contribute to semantic distinction between translated texts and original non-translated texts in the target language. Therefore, it is of great importance to test whether universals like simplification and levelling out influence the semantic features and informational structure of translated texts. Correspondingly, the involvement of parameters at the semantic level could provide valuable insights into the discussion of translation universals, and deepen our understanding of translation universals not only as syntactic phenomena but also as syntactic-semantic phenomena that are more complex and have a more profound impact on text characteristics at many different levels. This can also enhance cross-linguistic translation comparative studies and contribute to our understanding of translation as a complex system (Han & Jiang, 2017; Sang, 2023).

Regrettably, the exploration of translation universals from such a perspective is relatively sparse. This might be attributed to two major hurdles. One is the lack of automated semantic analytical methods for large-scale corpora. Despite the growth of corpus size, research in this area has proceeded for decades on manually created semantic resources, which has been labour-intensive and often confined to narrow domains (Màrquez et al., 2008). This deficiency has resulted in slow progress in the semantic analysis of translated texts. The other hurdle arises from the difficulty with extracting semantic features from texts across various corpora while minimizing the interference from different topics and content within these texts. The frequently-used techniques of deep semantic analysis, such as word vector models, are designed to capture word meanings, text theme, and context information, which makes them susceptible to the variance of textual content and thus unsuitable for comparing corpora consisting of both translated texts and non-translated original texts in the target language (Rong, 2014). To overcome these hurdles, the current study draws upon the insights from two natural language processing tasks and employs an approach driven by shallow semantic analysis, viz. semantic role labelling, and textual entailment analysis.

Specifically, two methods are adopted in the current study. They are respectively based on sentence-level semantic role labelling tasks and textual entailment tasks. They can facilitate the automation of the analysis without requiring too much context information and deep meaning. Additionally, semantic role labelling focuses on extracting the information structure of a sentence while textual entailment estimates the informational explicitness of a text. Since both methods perform semantic analysis without specifically considering word meaning and textual content, they are more suitable than deep semantic analysis tools for identifying the semantic universals of translated texts as well as distinguishing different language varieties.

Semantic role labeling and textual entailment. Semantic Role Labeling (SRL) is a Natural Language Processing (NLP) task designed to determine the precise semantic relations between a predicate and its associated participants and properties in a sentence. Its original theoretical base and annotation system are derived from the semantic roles and fundamental meaning relationships of case grammar (Fillmore, 1968).

Early attempts at SRL often rely on manual labelling and annotation. However, with advancements in linguistic theory, machine learning, and NLP techniques, especially the availability of large-scale training corpora (Shao et al., 2012), SRL tools have developed rapidly to suit technical and operational requirements. Nowadays, SRL models and tools boast high accuracy and robustness across different languages and domains, because they are based on theoretical achievements in phrase structure syntax and dependency syntax, together with deep learning models like long short-term memory networks and transformer architectures (Pradhan et al., 2005).

Three types of semantic roles are included in contemporary SRL annotation system: verbs that signify events, core arguments that represent the participants involved in the event (e.g. agents and

patients), and semantic adjuncts that describe other aspects of the event or participant relations (e.g. location and manner). A verb, together with one or more core arguments, forms the necessary semantic framework of a clause. Semantic adjuncts are seen as additional modifiers and determiners of the event (Xue & Palmer, 2009). By assigning semantic role labels to different elements in a sentence, SRL models reveal the syntactic-semantic structure underlying the sentence and provide a foundational semantic representation of the text, highlighting the fundamental event properties and relations among relevant entities expressed within the sentence. Compared with tools for syntactic annotation and analysis (e.g. dependency annotator) that put more emphasis on the role of prepositions and auxiliary words in dividing syntactic structures, SRL pays more attention to the semantic and logical relationship among content words (Che et al., 2021). Therefore, SRL offers a more comprehensive annotation that integrates both syntactic and semantic information from a sentence.

Recognizing Textual Entailment (RTE) is also an NLP task aimed at modelling language variability by identifying the textual entailment relationship between different words or phrases. Typically, RTE tasks involve two natural language expressions (mostly two sentences) that have a directional relationship. In these tasks, the entailing expression is referred to as the text (T), and the entailed expression is referred to as the hypothesis (H). A strict textual entailment can be detected when H can be inferred from T. That is to say, T contains the knowledge of H (Ferrández et al., 2006). The following example shows a true entailment between T1 and H1.

Example 1 An example of true entailment

T1	The sun rises in the east every morning.
H1	Sunrise occurs in the east.

Pazienza et al. (2005) proposed that three types of textual entailment can be distinguished operationally into semantic subsumption, syntactic subsumption, and direct implication. Semantic subsumption occurs when the Text presents the information more specifically than the Hypothesis through semantic operations. In the following example, T2 is semantically more specific than H2 due to the difference in the predicate used to describe the event:

Example 2 An example of semantic subsumption

T2	The cat devours the mouse.
H2	The cat eats the mouse.

Syntactic subsumption occurs when the information in the Text is presented more specifically than that in the Hypothesis through syntactic operations. For example:

Example 3 An example of syntactic subsumption

T3	The cat eats the mouse in the garden.
H3	The cat eats the mouse.

Direct implication refers to a situation in which the information expressed in the Hypothesis is inferred from the information in the Text. In the following example, H4 is implied by T4 even though the two predicates in them describe different events:

Example 4 An example of direct implication

T4	The cat eats the mouse.
H4	The cat killed the mouse.

In practical research, detecting direct implication requires the model to process deeper syntactic and semantic knowledge. Given this, the current study mainly focuses on semantic subsumption and syntactic subsumption, which can be readily captured through the analysis of relatively shallow semantic and syntactic information that is not overly deep and complex. Moreover, both semantic and syntactic subsumptions denote an exhaustive informational inclusion relationship between T and H, which means that T includes all the information in H, and H can be inferred from T. This indicates that the amount of information in T is equal to the amount of information in H plus extra information (E), which can be expressed as:

$$I(T) = I(H) + I(E) \quad (1)$$

The amount of extra information can also be interpreted as the distinction between implicit and explicit information, which can be captured through textual entailment. Take the semantic subsumption between T3 and H3 for example, I(E) is the information gap between the two predicates “eat” and “devour”. For the syntactic subsumption between T4 and H4, I(E) is the amount of information of the additional adverbial “in the garden”. Inspired by this idea, the current study attempts to compare the information explicitness in different corpora using methods based on semantic role labelling and textual entailment to examine whether translation universals such as explicitation and simplification exist at the syntactic-semantic level.

Specifically, the current study first divides the sentences in each corpus into different semantic roles. For each semantic role, a textual entailment analysis is then conducted to estimate and compare the average informational richness and explicitness in each corpus. Based on the results of textual entailment analysis, the study further investigates translation universals at the semantic level and collects evidence for the influence of the translation process on informational explicitness as well as the semantic structure.

In this study, we aim to answer the following research questions:

1. Do translation universals exist at the syntactic-semantic level? If so, what are the syntactic-semantic features typical of translated texts?
2. What factors contribute to the distinct features observed in translated texts at the syntactic-semantic level?

Methodology

Corpus. For a comprehensive understanding of S-universals and T-universals from a syntactic-semantic perspective, the current study uses English source texts, English-Chinese translations, and non-translated Chinese original texts (ES, CT, and CO, respectively) in two corpora as research objects. For the exploration of S-universals, ES are compared with CT in Yiyan English-Chinese Parallel Corpus (Yiyan Corpus) (Xu & Xu, 2021). Yiyan Corpus is a million-word balanced English-Chinese parallel corpus created according to the standard of the Brown Corpus. It contains 500 pairs of English-Chinese parallel texts of 4 genres with 1 million words in ES and 1.6 million Chinese characters in CT. For the exploration of T-universals, CT in Yiyan Corpus are compared with CO in the Lancaster Corpus of Mandarin Chinese (LCMC) (McEnery & Xiao, 2004). LCMC is a million-word balanced corpus of written non-translated original Mandarin Chinese texts, which was also created according to the standard of the Brown Corpus. Hence, it is comparable to the Chinese part of Yiyan Corpus in text quantity and genre. Overall, the research object of the current study is 500 pairs of parallel English-Chinese texts and 500 pairs of comparable CT and CO. All the raw

materials have been manually cleaned to meet the needs of annotation and data analysis.

Tools and research procedures. The semantic role labelling tools used for Chinese and English texts are respectively, Language Technology Platform (N-LTP) (Che et al., 2021) and AllenNLP (Gardner et al., 2018). N-LTP is an open-source neural language technology platform developed by the Research Center for Social Computing and Information Retrieval at Harbin Institute of Technology, Harbin, China. It offers tools for multiple Chinese natural language processing tasks like Chinese word segmentation, part-of-speech tagging, named entity recognition, dependency syntactic analysis, and semantic role tagging. N-LTP adopts the multi-task framework based on a shared pre-trained model, which has the advantage of capturing the shared knowledge across relevant Chinese tasks, thus obtaining state-of-the-art or competitive performance at high speed. (Che et al., 2021). AllenNLP, on the other hand, is a platform developed by Allen Institute for AI that offers multiple tools for accomplishing English natural language processing tasks. Its semantic role labelling model is based on BERT and boasts 86.49 test F1 on the Ontonotes 5.0 dataset (Shi & Lin, 2019).

In addition to a comprehensive analysis that includes all semantic roles, this study also focuses on several important roles to delve into the semantic discrepancies across the three text types. Considering the difference between Chinese and English semantic role tagsets, the current study chose some important and relatively frequent semantic roles as research focuses. The tagsets for both Chinese and English semantic role labelling of core arguments and semantic adjuncts are quite similar. Core arguments are labeled as ArgN or AN with N being numbers representing different types of relationships. For example, A0 represents the agent/causer/experiencer of the verb and A1 represents the patient and recipient of the verb. Semantic adjuncts are roles that are not directly related to the verb, typically determiners or roles that provide supplementary information about verbs and core arguments. Common semantic adjuncts include adverbials (ADV), manners (MNR), and discourse markers (DIS). The current study selects six of the most frequent semantic roles for in-depth investigation, including three core arguments (A0, A1, and A2) and three semantic adjuncts (ADV, MNR, and DIS).

After the semantic roles in each corpus are labelled, textual entailment analysis is then conducted based on the labelling results. For verbs, the analysis is mainly focused on their semantic subsumption since they are the roots of argument structures. For other semantic roles like locations and manners, the entailment analysis is mainly focused on their role in creating syntactic subsumption.

It should be noted that the textual entailment analysis employed in the current study introduces two modifications on the basis of typical RTE tasks, but the principle behind the two types of analysis remains the same, which is to analyze the semantic inclusion relationship between the text (T) and the hypothesis (H).

Firstly, typical RTE tasks determine whether there is an entailment relationship between T and H, but the textual entailment analysis employed in this study attempts to measure the distance or similarity between T and H when they form a determined entailment relationship. The distinctive aspect of our textual entailment analysis is that we take a given sentence as H and create its T by changing the predicate in the sentence into its root hypernym. In this way we manually create a determined entailment relationship between T and H. Based on this methodology, the extra information I(E) in Formula (1) can be approximated by the distance between the original predicate and its root hypernym. Then the distance can be quantified as 1 minus the Wu-Palmer Similarity or Lin Similarity between the original predicate and its root hypernym. In summary, Wu-Palmer Similarity or Lin Similarity actually provide a way to quantify and measure I(E) in Formula (1). By calculating the two values, we can approximate the explicit level of H to T, or in other words, the semantic depth of the original sentence H. A smaller the value of Wu-Palmer Similarity or Lin Similarity indicates a more explicit predicate.

Secondly, since the analysis of textual entailment involves a comparison between English and Chinese texts, multilingual semantic resources are needed. In the current study, the reference knowledge base for the textual entailment analysis in this study is WordNet (Miller, 1995) and its multilingual counterpart Open Multilingual WordNet (OMW). Numerous studies have proved that a shallow semantic analysis based on WordNet is adequate for monolingual and multilingual RTE tasks (Castillo, 2011; Ferrández et al., 2006; Reshmi & Shreelekshmi, 2019).

The current study uses several syntactic-semantic features as indices to represent the syntactic-semantic features of each corpus from the perspective of syntactic and semantic subsumptions. For syntactic subsumption, all semantic roles are described with features across three dimensions, viz. average number of semantic roles per verb (ANPV), average number of semantic roles per sentence (ANPS), and average role length (AL). ANPV and ANPS reflect syntactic complexity and semantic richness respectively in clauses and sentences. Compared to measurements using purely syntactic components, such measurements focusing on semantic roles can better indicate substantial changes in information quantity. AL reflects the information quantity within a semantic role. These indices are intended to detect information gaps resulting from syntactic subsumption, which often takes the form of either an increase in number of semantic roles or an increase in the length of a single semantic role.

For semantic subsumption, verbs that serve as the roots of argument structures are evaluated based on their semantic depth, which is assessed through a textual entailment analysis based on WordNet. The identification of semantic similarity or distance between two words mainly relies on WordNet's subsumption hierarchy (hyponymy and hypernymy) (Budanitsky & Hirst, 2006; Reshmi & Shreelekshmi, 2019). Therefore, each verb is compared with its root hypernym and the semantic distance between them can be interpreted as the explicitness of the verb. A bigger distance between a verb and its root hypernym indicates a deeper semantic depth and a higher level of explicitness. The WordNet module in the Natural Language Toolkit (NLTK) includes some measures previously developed to quantify the semantic distance between two words. Some of them are computed over semantic networks while others are combined with the notion of Information Content (IC) from information theory. Therefore, the current study chose Wu-Palmer Similarity and Lin Similarity as the measures employed in the analysis to include both types of measures.

Wu-Palmer Similarity (Wup Sim) was first introduced as a conceptual similarity that measures the similarity between two-word senses (s_1 and s_2) by considering the depth of both senses and the depth of their least common subsumer (*lcs*) in the taxonomy (Wu & Palmer, 1994). Its calculation is completely dependent on the relationships and paths in the semantic network. It can be calculated as below:

$$\mathrm{sim}\left(\mathrm{s}_1, \mathrm{s}_2\right) = \frac{2 \times \mathrm{D}}{\mathrm{L}_1 + \mathrm{L}_2 + 2 \times \mathrm{D}} \tag{2}$$

in which L1 and L2 represent, respectively, the path length between *lcs* and s_1, s_2 while D represents the depth of *lcs*.

The value range of values for Wu Palmer Similarity is [0, 1], where 0 indicates dissimilar and 1 indicates completely similar.

Lin Similarity (Lin Sim) is also known as Lin's Universal Similarity Measure which is applicable to arbitrary objects without presuming any form of knowledge representation (Lin, 1998). It measures the similarity between s_1 and s_2 based on their information content (IC) as well as the information content of their *lcs*. Lin Similarity can be calculated as below:

$$\mathrm{sim}\left(s_1, s_2\right) = \frac{2 \times \mathrm{IC}(lcs)}{\mathrm{IC}\left(s_1\right) + \mathrm{IC}\left(s_2\right)} \quad (3)$$

In the current study, the information content is obtained from the Brown information content database (ic-brown.dat) integrated into NLTK. Like Wu-Palmer Similarity, Lin Similarity also has a value range of [0, 1], where 0 indicates dissimilar and 1 indicates completely similar.

Results

S-universals. This section mainly focuses on the discussion of S-universals and presents the results of the comparison between ES and CT. With all the data collected, several statistical tests were conducted on all the indices to explore whether CT exhibit significant semantic differences from ES. Then, a detailed inspection of specific semantic roles was conducted to discuss specific semantic divergences between the two text types.

To begin with, Leneve's tests were conducted on each index to see whether there was a homogeneity of variance. The results in Table 1 indicate that there are unequal variances between ES and CT for all indices. Plus, the distributions of some semantic features do not exhibit normality. Thus, several Mann-Whitney U tests were performed to determine whether there are significant differences between the indices of the two different text types.

In Table 2, the five indices and the results of the Mann-Whitney U tests indicate that there is a notable divergence between CT and ES, with significant differences for most indices.

Semantic subsumption. In terms of semantic subsumption, the results of both Wu-Palmer Similarity and Lin Similarity in Table 2 indicate that verbs in CT are less similar to their root hypernyms than those in ES. As a result, they seem to have a deeper average semantic depth and a higher level of explicitness than verbs in ES. The results of Mann-Whitney U tests indicate statistically significant results, implying that verbs in CT show a quite pronounced characteristic of explicitation in terms of semantic subsumption.

A closer inspection of the entailment analysis results revealed a substantial diversity between Chinese verbs and English verbs that could account for the significant difference in semantic subsumption. English sentences use "be" verbs (is, are, etc.) much more frequently, whose Wu-Palmer Similarity and Lin Similarity values are both 1. However, the frequency of their Chinese corresponding verbs, such as "是*(is/are)*" in CT, is notably lower. Instead, the "be" verbs functioning as predicates in ES are often substituted in CT with other notional verbs, which contributes greatly to the lower average Wu-Palmer Similarity and Lin Similarity of CT. For example:

Example 5 (Text Pair A08, Sentence 27)

Source text:	**Since then it has been a steady slide, to a low of 25 percent just prior to the election.**
Translation:	自 那时 起 该 支持率 一路 下滑 , 到 大选 From then begin this rate of support all the way decline , to election 前 只有 25% 。 before only 25% .

In the above example, the verb in the source text is "*been*", but the predicate is changed to the verb "下滑(*decline*)" in the translation, which comes from the word "*slide*" in the source text. Transformation in predicates of this kind, known as denominalization, is essentially one of the major factors contributing to the difference in semantic depths of verbs. According to Systemic Functional Linguistics theory, nominalization illustrated in the source text causes an incongruent or metaphorical relationship between the lexico-grammar layer and the semantic layer in the stratal model (Halliday, 1985; Halliday, 1993; Halliday & Martin, 1993), which leads to grammatical metaphor (Halliday, 1985; 1993; Halliday & Matthiessen, 2006; Taverniers, 2006) and makes the information more concise but less explicit (McGrath & Liardét, 2023))e.g. the meaning of "*decline*" is implied by the noun "*slide*". Through denominalization in the translation process, the notion of "*decline*" is reintroduced to the predicate verb, which eliminates the incongruency between the lexico-grammatical and semantic layers, resulting in more explicit information. To sum up, the semantic subsumption analysis not only reveals that verbs in CT exhibit a higher level of explicitness than verbs in ES, but it also pinpoints a major cause for this significant difference, namely the transformation of the information structure at the sentence level, which is achieved through denominalization in the translation process.

Syntactic subsumption. Table 2 shows that the average number of semantic roles per sentence (ANPS) of CT is approximately the same as that of ES. However, CT's average number of semantic roles per verb (ANPV) and average role length (ARL) are significantly lower than those of ES. This suggests that argument

Table 1 Leneve's tests on syntactic-semantic features of ES and CT.

	F	df1	df2	*p*
Wup Sim	5.18	1	998	0.023
Lin Sim	18.50	1	998	<0.001
ANPV	11.67	1	998	0.001
ANPS	4.22	1	998	0.040
AL	208.74	1	998	<0.001

Table 2 Mann-Whitney U tests on overall syntactic-semantic features of ES and CT.

		ES		CT		Mann-Whitney U tests	
		mean	std.	mean	std.	*Z*	*p*
Semantic Subsumption Features	Wup Sim	0.66	0.04	0.50	0.04	−27.13	<0.001
	Lin Sim	0.83	0.03	0.78	0.03	−22.59	<0.001
Syntactic Subsumption Features	ANPV	2.64	0.16	2.17	0.13	−26.64	<0.001
	ANPS	7.75	1.83	7.75	2.05	−0.45	0.651
	ARL	4.45	1.00	2.77	0.50	−24.05	<0.001

Table 3 Mann-Whitney U tests on syntactic-semantic features of specific semantic roles in ES and CT.

		ES		CT		Mann-Whitney U tests	
		mean	std.	mean	std.	*Z*	*p*
ANPV	A0	0.52	0.10	0.59	0.09	−9.38	<0.001
	A1	0.92	0.05	0.68	0.05	−27.34	<0.001
	A2	0.33	0.05	0.07	0.02	−27.37	<0.001
	ADV	0.11	0.03	0.49	0.09	−27.37	<0.001
	MNR	0.09	0.03	0.02	0.02	−25.78	<0.001
	DIS	0.05	0.02	0.12	0.04	−24.05	<0.001
	Overall	2.64	0.16	2.17	0.13	−26.64	<0.001
ANPS	A0	1.55	0.46	2.08	0.59	−14.56	<0.001
	A1	2.70	0.58	2.39	0.58	−8.40	<0.001
	A2	0.98	0.25	0.24	0.09	−27.33	<0.001
	ADV	0.32	0.13	1.76	0.60	−27.34	<0.001
	MNR	0.25	0.12	0.08	0.06	−24.20	<0.001
	DIS	0.14	0.08	0.45	0.22	−24.03	<0.001
	Overall	7.75	1.83	7.75	2.05	−0.45	0.65
ARL	A0	2.47	0.99	2.17	0.57	−3.32	<0.001
	A1	5.38	1.34	4.20	0.91	−14.78	<0.001
	A2	6.20	1.74	3.98	1.27	−19.48	<0.001
	ADV	6.02	1.75	1.17	0.10	−27.37	<0.001
	MNR	4.20	1.44	4.48	1.52	−3.28	<0.001
	DIS	1.32	0.31	1.13	0.18	−11.71	<0.001
	Overall	4.45	1.00	2.77	0.50	−24.05	<0.001

structures in CT normally contain semantic roles that are fewer and shorter than those in ES. In terms of syntactic subsumption, it seems that CT have an inclination for simplification in argument structure. Moreover, the average number of argument structures in Chinese sentences should be bigger than that in English sentences since they have a similar average number of semantic roles in a sentence. In other words, the results of syntactic subsumption analysis indicate an "unpacking" process from ES into CT, during which relatively long semantic roles in English sentences are simplified and broken down into shorter roles, or even transformed into several new argument structures, thus resulting in shortened average role length and simplified argument structures.

It should be noted that the significant difference in ARL could potentially be ascribed to linguistic diversity between Chinese and English (e.g. more frequent functional words in English texts) instead of syntactic subsumption. To address this issue, this study standardized ARL with sentence length and tested if there was a significant difference between their proportions in sentences to test if ARL reflects semantic differences. The standardized ARLs of English and Chinese semantic roles are respectively 0.14 and 0.09. The Mann-Whitney U tests show that there is also a significant difference between them ($Z = -24.79$, $p < 0.001$). This corroborates the presence of syntactic subsumption between CT and ES in the difference in ARL.

For a more detailed view of the differences in syntactic subsumption between CT and ES, the current study analyzed the features of several important semantic roles. The results of the comparison between each role are shown in Table 3.

Table 3 indicates that significant differences between CT and ES can be observed in almost all the features of the semantic roles. For core arguments that are the main components constituting the semantic structure of a sentence, the differences in all the features add weight to the proposition that information structures of sentences in CT exhibit characteristics substantially different from those in ES for several reasons. First, the values of ANPV and ANPS of agents (A0) in CT are significantly higher than those in ES, suggesting that Chinese argument structures and sentences usually contain more agents. This could serve as evidence for translation explicitation, in which the translator adds the originally omitted sentence subject to the translation and make the subject-verb relationship explicit. On the other hand, all the syntactic subsumption features (ANPV, ANPS, and ARL) for A1 and A2 in CT are significantly lower in value than those in ES. Consequently, these two roles are found to be shorter and less frequent in both argument structures and sentences in CT, which is in line with the above-assumed "unpacking" process.

As for semantic adjuncts, it is worth noting that the average number of discourse markers (DIS) in CT is significantly bigger than that in ES, indicative of the translator's inclination to enhance the coherence and thus the necessity to make certain contextual logical relationships explicit. Additionally, the number of adverbials (ADV) in CT is significantly bigger than that in ES while the number of manners (MNR) in CT is significantly smaller. With both semantic roles being modifiers of verbs, this finding reconfirms our hypothesis that the English-Chinese translation process has a denominalizing effect since some of the MNR in English source texts are converted (e.g. "do sth like/as…" or "do sth in the manner of…") into adverbial modifiers.

Following is an example illustrating the transformation of sentence-level information structure:

Example 6 (Text Pair J51 Sentence 25)

Source text:	**It makes us forget our potential for naturalness, which, for all its uncertainty, is more of a clue to our future than the certainty our abstract knowledge gives us.**
Translation:	它 使 我们 忘记 了 我们 在 自然 本性 上 的 It make us forget we natural character 潜能 。 由于 这种 潜能 的 不确定性 , 它 potential . Because of this potential uncertainty , it 只 是 我们 未来 的 线索 , 而 不 是 我们的 only is our future clue , yet not is our 抽象 知识 给予 我们 的 确定性 。 abstract knowledge give us certainty .

In the above example, an English compound sentence is divided and translated into two Chinese sentences, whose results of semantic role labeling are shown in Figs. 1 and 2.

With all the argument structures in the above example compared, two major effects of the divide translation can be found in the features of semantic roles. The shortened role length is the first and most obvious effect, especially for A1 and A2. In the English sentence, the longest semantic role contains 27 words while the longest role in Chinese sentences contains only 9 words. As can be readily seen in Fig. 1, extremely long roles can be attributed to multiple substructures nested within the semantic role, such as A1 in Structure 1 (Fig. 1) in the English sentence, which contains three sub-structures. According to the cognitive load theory (Sweller, 2011), this multi-layered nested structure forces the readers to store the information of all the upper layers in memory while processing information from the bottom layer, which contributes significantly to their cognitive load. In contrast, this multi-layered nested structure is deconstructed and decomposed in translated texts through the divide translation, and the number of sub-structures contained in each semantic role is controlled no greater than 1. This example proves that the informational structures in the translated texts are significantly simplified by reducing the number of nested sub-structures in semantic roles.

The other major effect lies in the conversion and addition of certain semantic roles for logical explicitation. In Structure 3 (Fig. 2), the Chinese translation converted the role of adverbial (ADV) in the source text into a purpose or reason (PRP) by adding the specific logical symbol "由于*(because of)*". Also, the discourse marker "而 *(yet)* " is added in Structure 2 (Fig. 2). These instances of conversion and addition are essentially a shift from logical grammatical metaphors to congruent forms that occurs

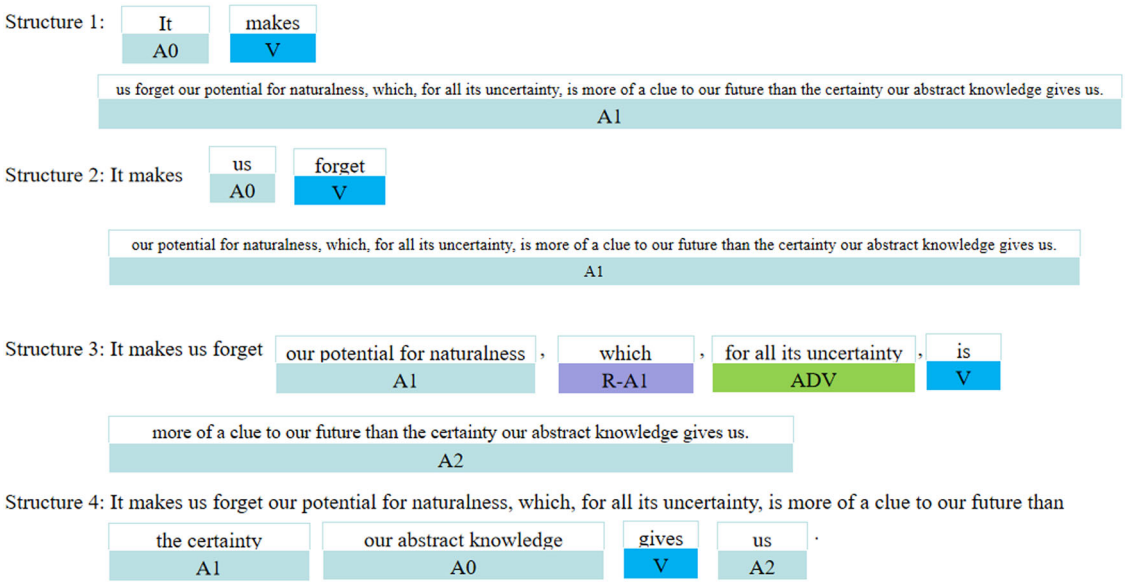

Fig. 1 Results of semantic labelling of the source text in Example 6. There are altogether 4 argument structures nested in the English sentence, with each semantic role in the structure highlighted and labelled. The hierarchical nestification structure is illustrated by the fact that one sub-structure functions as a semantic role (usually A1 or A2) in its dominative argument structure.

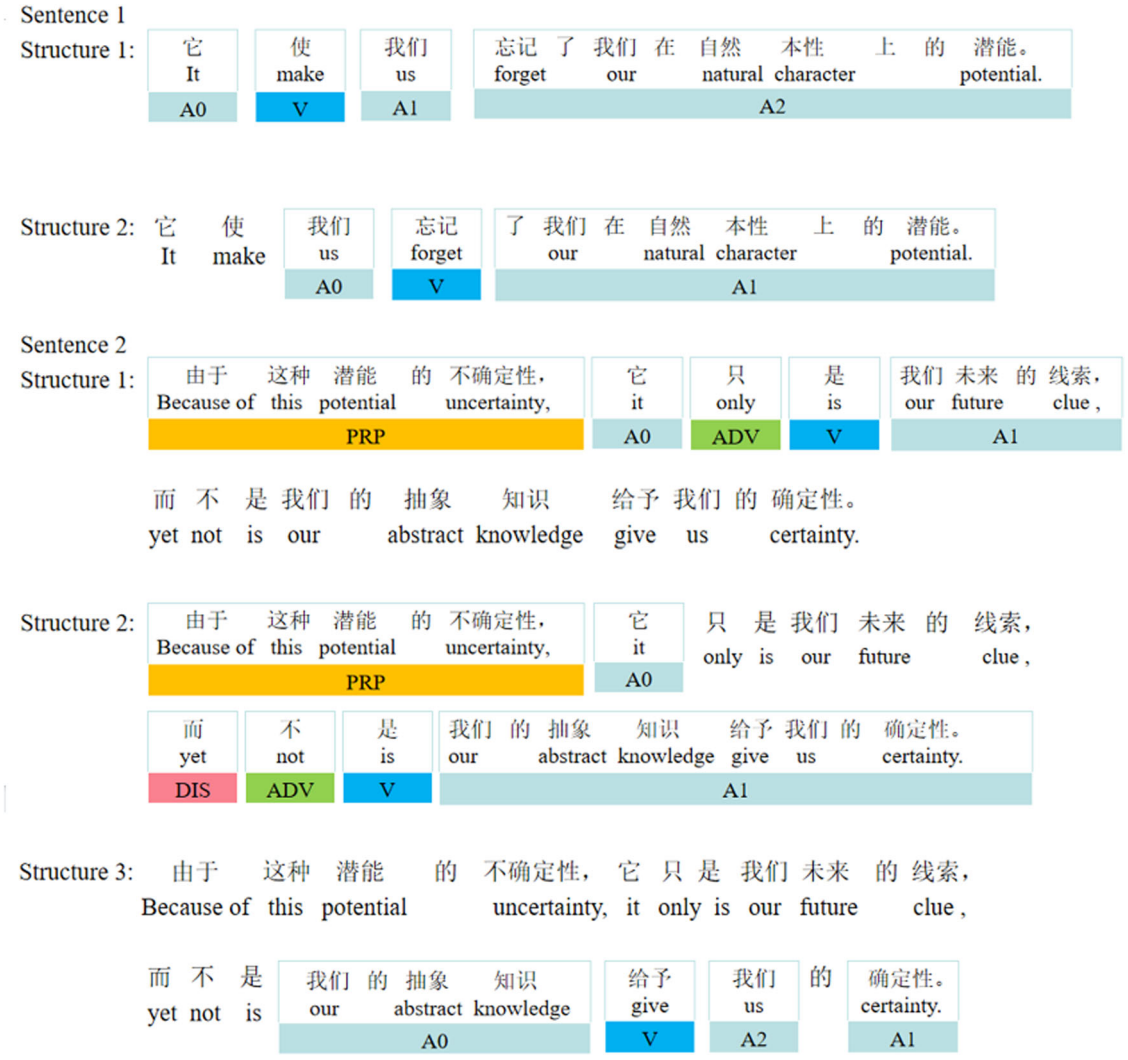

Fig. 2 Results of semantic role labelling of the translation in Example 6. The original English sentence is split into two Chinese sentences through divide translation. Sentence 1 contains a two-layered hierarchical nestification structure while Sentence 2 contains a three-layered hierarchical nestification structure. Each semantic role in the structure is highlighted and labelled.

during the translation process, through which the logical semantic is made explicit (Martin, 1992).

In summary, the analysis of semantic and syntactic subsumptions reveals many significant divergences between ES and CT at the syntactic-semantic level. For specific S-universals, some evidence for explicitation is found in CT, such as a higher level of explicitness for verbs and a higher frequency of agents (A0) and discourse markers (DIS). Evidence for simplification in information structure is also found in the form of fewer syntactic nestifications, illustrated mainly by a shorter role length of patients (A1) and ranges (A2). Based on these divergences, it is safe to conclude that CT do show a syntactic-semantic characteristic significantly distinct from ES.

T-universals. This section focuses on T-universals and presents the results of the comparison between CT and CO. The results of Leneve's tests in Table 4 exhibit unequal variances between CO and CT for all indices. Mann-Whitney U tests were then conducted to determine whether there were significant differences in indices between two different text types.

Semantic subsumption. Table 4 shows that CT exhibit average Wu-Palmer Similarity and Lin Similarity values notably similar to those of CO, which is logically consistent as both text types operate within the same language system, inherently sharing linguistic characteristics. Although the differences are still statistically significant with small p values, the effect size of the U test on Lin Similarity is only 0.092, which is not big enough to support a significant effect. Thus, other methods must be employed to further determine whether there is a noticeable difference in semantic subsumption between CT and CO.

Table 4 Leneve's tests on syntactic-semantic features of CO and CT.

	F	df1	df2	p
Wup Sim	4.43	1	998	0.036
Lin Sim	32.60	1	998	<0.001
ANPV	4.71	1	998	0.030
ANPS	35.96	1	998	<0.001
ARL	5.18	1	998	0.023

To have a better understanding of the nuances in semantic subsumption, this study inspected the distribution of Wu-Palmer Similarity and Lin Similarity of the two text types. The results of the inspection are illustrated in Figs. 3 and 4.

The two figures show that while the two text types exhibit similar average values of Wu-Palmer Similarity and Lin Similarity, differences can still be observed in their distributions, with more translated texts concentrated at a relatively higher level compared to non-translated texts, most of which register at a relatively lower level of average Wu-Palmer Similarity and average Lin Similarity. Therefore, the difference in semantic subsumption between CT and CO does exist in the distribution of semantic depth. On the one hand, U test results indicate a generally higher level of explicitation in verbs of CO than those of CT. On the other hand, the comparison of the distributions reveals that semantic subsumption features of CT are more centralized than those of CO, which can be understood as a piece of evidence for levelling out.

Levelling out, as one of the sub-hypotheses of translation universals, is defined as the inclination of translations to "gravitate towards the center of a continuum" (Baker, 1996). It is also called "convergence" by Laviosa (2002) to suggest "the relatively higher level of homogeneity of translated texts". Under the premise that the two corpora are comparable, the more centralized distribution of translated texts indicates that semantic subsumption features of CT are relatively more consistent than the higher variability of CO.

Syntactic subsumption. Table 5 shows that translated texts' syntactic subsumption features of CT are higher than those of CO. This suggests that in CT, argument structures and sentences typically feature more and longer semantic roles than in CO. From these results we can infer that sentences in CT may have a more complex and condensed syntactic-semantic structure with a higher density of semantic roles in argument structures as well as sentences than in CO.

In our further exploration of specific semantic roles, results of the Mann-Whitney U tests in Table 6 show that there exist significant differences in most features across various semantic roles, suggesting that CT are quite distinct from CO in syntactic-semantic strictures.

For semantic adjuncts, the results show that the p-values of the comparison between the ANPS of adverbials (ADV) and manners (MNR) are smaller than 0.05. However, the effect sizes of the two

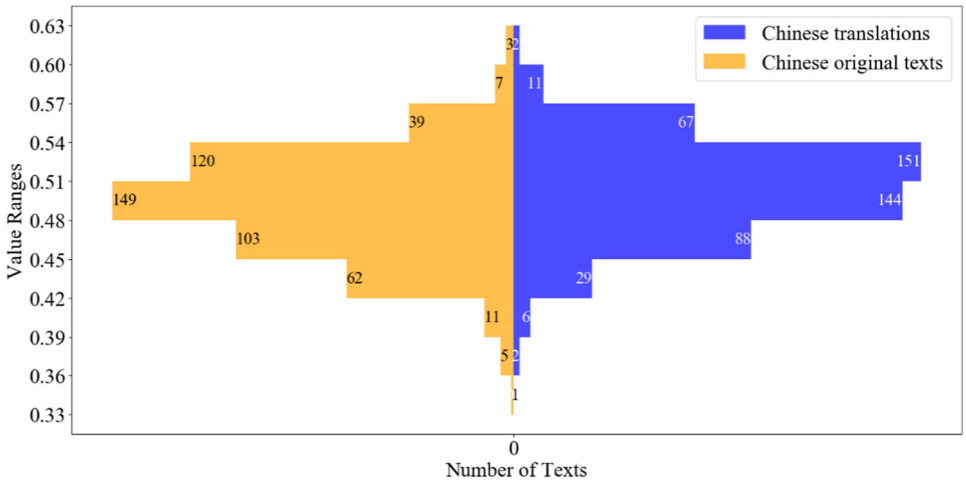

Fig. 3 Wu-Palmer Similarity distributions of CT and CO. The value range of Wu-Palmer Similarity is divided into 10 subintervals, and the number of texts in CT and CO that fall into each subinterval is counted. This figure provides a clearer illustration of the nuanced differences between the Wu-Palmer Similarity distributions of CT and CO than a boxplot.

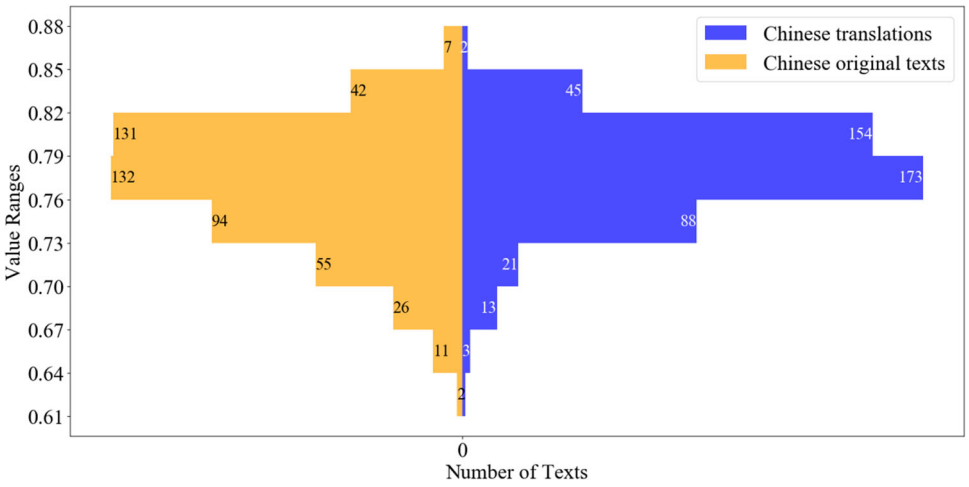

Fig. 4 Lin Similarity distributions of CT and CO. The value range of Lin Similarity is divided into 9 subintervals, and the number of texts in CT and CO that fall into each subinterval is counted. This figure provides a clearer illustration of the nuanced differences between the Lin Similarity distributions of CT and CO than a boxplot.

Table 5 Mann-Whitney U tests on overall syntactic-semantic features of CO and CT.

		CO		CT		Mann-Whitney U tests	
		mean	std.	mean	std.	Z	p
Semantic subsumption	Wup Sim	0.49	0.04	0.50	0.04	−4.68	<0.001
	Lin Sim	0.77	0.04	0.78	0.03	−2.87	0.004
Syntactic subsumption	ANPV	2.09	0.14	2.17	0.13	−9.97	<0.001
	ANPS	6.92	1.54	7.75	2.05	−6.85	<0.001
	ARL	2.67	0.56	2.77	0.50	−3.33	0.001

Table 6 Mann-Whitney U tests on syntactic-semantic features of specific semantic roles in CO and CT.

		CO		CT		Mann-Whitney U tests	
		mean	std.	mean	std.	*Z*	*p*
ANPV	A0	0.53	0.10	0.59	0.09	−8.50	<0.001
	A1	0.66	0.07	0.68	0.05	−4.40	<0.001
	A2	0.06	0.02	0.07	0.02	−6.71	<0.001
	ADV	0.50	0.10	0.49	0.09	−0.90	0.37
	MNR	0.03	0.03	0.02	0.02	−4.03	<0.001
	DIS	0.10	0.05	0.12	0.04	−6.69	<0.001
	Overall	2.09	0.14	2.17	0.13	−9.97	<0.001
ANPS	A0	1.75	0.44	2.08	0.59	−9.78	<0.001
	A1	2.18	0.51	2.39	0.58	−6.20	<0.001
	A2	0.20	0.10	0.24	0.09	−8.29	<0.001
	ADV	1.66	0.54	1.76	0.60	−2.60	0.01
	MNR	0.10	0.08	0.08	0.06	−2.70	0.01
	DIS	0.35	0.19	0.45	0.22	−7.54	<0.001
	Overall	6.92	1.54	7.75	2.05	−6.85	<0.001
ARL	A0	2.32	0.68	2.17	0.57	−3.22	<0.001
	A1	3.81	0.82	4.20	0.91	−7.32	<0.001
	A2	3.73	1.57	3.98	1.27	−4.56	<0.001
	ADV	1.16	0.11	1.17	0.10	−3.91	<0.001
	MNR	4.48	1.90	4.48	1.52	−0.87	0.39
	DIS	1.21	0.25	1.13	0.18	−5.59	<0.001
	Overall	2.67	0.56	2.77	0.50	−3.33	<0.001

U tests are not big enough (relatively 0.083 and 0.086) to support significant differences. On the other hand, ANPS of discourse markers (DIS) in CT is significantly higher than that in CO with a relatively larger effect size (0.241), indicating a higher frequency of discourse markers in CT.

For core arguments, the results show that the syntactic-semantic structures of CT are more complex than those of CO, with ANPV and ANPS of all the core arguments being significantly higher. Given the comparison between CT and ES, this could result from "the source language shining-through hypothesis", which is defined as the source language's interference with the translation process (Teich, 2003). It can cause the translation to retain some of the lexical and grammatical features of the source language (Dai & Xiao, 2010; Xiao, 2015). As discussed in previous sections, syntactic-semantic structures in ES have significant complexity characterized by nominalization and syntactic nestification. Although most syntactic-semantic structures are simplified through denominalization and divide translation in the translation process, a small portion of the sentences in CT retain the features of syntactic subsumption of ES. This results in the fact that CT exhibit traits that are unique to CO.

Discussion

Based on the above results, it can be concluded that CT do show several distinctions from both ES and CO at the syntactic-semantic level, which can be evidenced by the significant differences in syntactic-semantic features. These distinctions partially support the hypotheses of "the third language" and some translation universals.

For specific sub-hypotheses, explicitation, simplification, and levelling out are found in the aspects of semantic subsumption and syntactic subsumption. However, it is worth noting that syntactic-semantic features of CT show an "eclectic" characteristic and yield contrary results as S-universals and T-universals. For example, the average role length of CT is shorter than that of ES, exhibiting S-simplification. But the average role length of CT is longer than that of CO, exhibiting T-sophistication. This contradiction between S-universals and T-universals suggests that translation seems to occupy an intermediate location between the source language and the target language in terms of syntactic-semantic characteristics. This finding is consistent with Fan and Jiang's (2019) research in which they differentiated translational language from native language using mean dependency distances and dependency direction. They found syntactic eclectic features of translated texts at the syntactic level, suggesting that translation is the result of the negotiation between the source language and the target language, liable to influences from both directions (Fan & Jiang, 2019). In the current study, such eclectic features are also found at the syntactic-semantic level, indicating that the negotiation in the complex translation process also has an impact on the semantic characteristic of the translated texts. This supports Krüger's (2014) view that S-universal and T-universal are caused by different factors. One plausible explanation for these findings might be the Hypothesis of Gravitational Pull posited by Halverson (2003, 2017), which assumes that translated language is affected by three types of forces. One force is the "magnetism effect" of the target language that comes from prototypical or highly salient linguistic forms. The second force is the "gravitational pull effect" that comes from the source language, which is the counter force of the magnetism effect that stretches the distance between the translated language and the target language. The third force comes from the "connectivity effect" that results from high-frequency co-occurrences of translation equivalents in the source and the target languages (Halverson, 2017). This hypothesis, which has been used to explain translation universals at the lexical and syntactic levels (Liu et al., 2022; Tirkkonen-Condit, 2004) may also extend its applicability to translation universals at the semantic level. The results of the current study suggest that the influences of both the source and the target languages on the translated language are not solely limited to the lexical and syntactic levels. Notably, these influences also manifest distinctly at the semantic level.

Specifically, on the one hand, the target language's "magnetism effect" can be substantiated by denominalization and divide translation, as discussed in the previous section. On the other hand, examples of the "gravitational pull effect" and the "connectivity effect" can also be found to cause the diversity between CT and CO. For example, the connectivity effect can lead to differences in semantic subsumption, as demonstrated by the following example,

Example 7 (Text Pair A02 Sentence 82)

Source text:	**Our expectation is that we would be able to travel and engage with the Chinese as soon as possible.**								
Translation:	我们的	期望	是	能	尽可能早地	成行	与	中国	
	Our	expectation	is	be able to	as soon as possible	travel	with	China	
	洽谈	。							
	negotiate	.							

In this example, the contextual need for de-nominalization is overshadowed by the "connectivity effect", causing the translation to retain the nominalization and the predicate "is" from the source text. This leads to an idiosyncratic information structure in the target language and hence, the deviation between the translated and target languages.

In terms of syntactic subsumption, the "gravitational pull effect" can be illustrated by the following example.

Example 8 (Text Pair F14 Sentence 40)

Source text:	**I think marriage takes really talented dreamers and creative beings that are capable of creating real change and puts them inside this widely accepted institution of marriage...**											
Translation:	我	认为	婚姻	需要	那些	有	能力	创造	真正的			
	I	think	marriage	need	those	have	ability	create	real			
	变化	并	把	它们	放进	被	普遍	接受	的	婚姻		
	change	and		them	put in		widely	accept		marriage		
	制度	里	的	真正	有	天赋	的	梦想家	和	创造者	…	
	system			real	have	talent		dreamer	and	creator	...	

In the above example, the translation follows the information structure of the source text and retains the long attribute instead of dividing it into another clause structure. The result is a massive nestification of a five-layered argument structure with a high degree of complexity, a feature that rarely manifests in the target language. This demonstrates how deviation between the translated language and target language is generated under the influence of the source language, also referred to as the "source language shining through" (Dai & Xiao, 2010; Teich, 2003; Xiao, 2015).

Overall, the Hypothesis of Gravitational Pull provides a framework for explaining the eclectic characteristics of syntactic-semantic features in the translated texts. The results of the current study support the hypothesis that syntactic-semantic features of translations are shaped by an equilibrium across the counteracting forces of the "magnetism effect", the "gravitational pull effect" and the "connectivity effect" (Halverson, 2003, 2017). This results in a distinct syntactic-semantic characteristic of translations that may deviate from both source and target languages, hence an eclecticism.

However, intriguingly, some features of specific semantic roles show characteristics that are common to both S-universal and T-universal. For example, the frequencies of agents (A0) and discourse markers (DIS) in CT are higher than those in both ES and CO, suggesting that the explicitation in these two roles is both S-oriented and T-oriented. This indicates that while syntactic-semantic features of translations are influenced by source and target language systems, they can also be driven by various other factors (e.g. translation norms and socio-cultural factors) and exhibit distinct characteristics that are beyond the source and target languages (Bernardini & Ferraresi, 2011; Muñoz Martín & Martín de León, 2020; Pym, 2005; Toury, 1995). In other words, there is an additional force that drives the translated language away from both the source and target language systems, and this force could be pivotal in shaping translated language as "the third language" or "the third code".

That is to say, translation universals at the syntactic-semantic level, such as explicitation and simplification, can be further distinguished depending on whether the syntactic-semantic feature presents the same or opposite results for S-universal and T-universal. This further suggests that even the translation universal under the same sub-hypothesis, like explicitation as S-universal, can be attributed to different causes. In this study, some cases of semantic explicitation, illustrated by de-nominalization (e.g. Example 4), can be attributed to the magnetism effect of the target language, while other cases of explicitation, illustrated by higher frequencies of agents and discourse markers, are more likely to be attributed to an additional force, which can be assumed as socio-cultural factors or the translator's factors (e.g., the translator may make the information clearer and more explicit to manage the risk of non-cooperation in the communication) (Pym, 2005). Therefore, further analysis is

warranted to distinguish different types of translation universals at the syntactic-semantic level and figure out the underlying causes so that we can better understand translation as a dynamic and complex system (Han & Jiang, 2017; Sang, 2023).

Conclusion

Using semantic role labeling and textual entailment analysis, the current study compared Chinese translations (CT) across English source texts (ES) and non-translated Chinese original texts (CO) to determine whether translation universals exist at the syntactic-semantic level. Investigations on semantic subsumption and syntactic subsumption in both S-universals and T-universals have found significant differences across the three text types, suggesting that CT do deviate significantly from ES as a parallel corpus and from CO as a comparable corpus as well. Substantial evidence for syntactic-semantic explicitation, simplification, and levelling out is found in CT, validating that translation universals are found not only at the lexical and grammatical levels but also at the syntactic-semantic level. Notably, the results indicate that overall syntactic-semantic features of CT exhibit an "eclectic" characteristic represented by contrary results for S-universal and T-universal, which could be attributed to the influence of both the source language and the target language, suggesting that S-universal and T-universal are cause by forces from different directions. On the other hand, explicitations are also found consistently as both S-universal and T-universal for certain specific semantic roles (A0 and DIS), which reflects the influence of socio-cultural factors in addition to the impact of language systems. These findings have further proved that translation is a complex system formed by the interplay of multiple factors (Han & Jiang, 2017; Sang, 2023), resulting in the diversity and uniqueness of translated language.

Limitations and future research directions

It should be acknowledged that although semantic role labeling and textual entailment analysis in this study provide some insights into the syntactic-semantic distinction of Chinese translations from English source texts and non-translated Chinese original texts, its findings serve as initial insights rather than conclusive findings about translation universals since they are limited to only one language pair. Further studies are needed to explore whether similar distinction exists in other language pairs, especially those having a higher level of similarity in information structures.

The discussion regarding the interaction between different semantic roles within an argument structure is limited in this study since the interaction process is not the primary variable of focus and the indices are designed to reflect the characteristics of the entire text group instead of sentence-level features. Nevertheless, an exploration of the interaction between different semantic roles is important for understanding variations in semantic structure and the complexity of argument structures. Hence, further studies are encouraged to delve into sentence-level dynamic exploration of how different semantic elements interact within argument structures.

Furthermore, many details in the research process have much room for further improvement. Additional features, such as indices for contextual semantic characteristics and the number of argument structure nestifications, could be included in the analysis. Moreover, the current study does not involve the refinement of semantic analysis tools since the modification and improvement of language models require high technique level and a massive quantity of training materials. Nonetheless, it is imperative for further studies to enhance these models and tools for semantic labelling and analysis, so as to promote a deeper understanding of semantic structures across different text types and languages.

Data availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Received: 19 January 2024; Accepted: 10 June 2024;
Published online: 27 June 2024

References

Baker M (1993) Corpus linguistics and translation studies—implications and applications. In Text and technology. John Benjamins
Baker M (1996) Corpus-based translation studies: The challenges that lie ahead. Terminology, LSP, and Translation: Studies in Language Engineering in Honour of Juan C. Sager 18:175
Bernardini S, Ferraresi A (2011) Practice, description and theory come together–normalization or interference in italian technical translation? Meta 56(2):226–246
Blum-Kulka S (1986) Shifts of cohesion and coherence in translation. Interling. Intercult Commun.: Discourse Cogn Ttransl Second Lang. Acquis. Stud. 272:17
Budanitsky A, Hirst G (2006) Evaluating wordnet-based measures of lexical semantic relatedness. Comput Linguist. 32(1):13–47
Castillo JJ (2011) A wordnet-based semantic approach to textual entailment and cross-lingual textual entailment. Int. J. Mach. Learn. Cybern. 2:177–189
Che W, Feng Y, Qin L, Liu T (2021) N-ltp: An open-source neural language technology platform for Chinese. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations
Chesterman A (2004) Beyond the particular. In Mauranen A & Kujamaki P (Eds.), Translation universals. Do they exist? John Benjamins
Dai G, Xiao R (2010) 'Sl shining through'in translational language: A corpus-based study of chinese translation of english. Proceedings of The International Symposium on Using Corpora in Contrastive and Translation Studies 2010 Conference (UCCTS2010), Lancaster University
Duff A (1981) The third language: Recurrent problems of translation into english: It ain't what you do, it's the way you do it. Pergamon Press
Eskola S (2004) Untypical frequencies in translated language: A corpus-based study on a literary corpus of translated and non-translated finnish. In Mauranen A & Kujamaki P (Eds.), Translation universals: Do they exist? John Benjamins
Fan L, Jiang Y (2019) Can dependency distance and direction be used to differentiate translational language from native language? Lingua 224:51–59
Ferrández Ó, Terol RM, Munoz R, Martínez-Barco P, Palomar M (2006) Deep vs. Shallow semantic analysis applied to textual entailment recognition. International Conference on Natural Language Processing, Finland
Fillmore CJ (1968) Lexical entries for verbs. Foundations of language, 373-393
Frawley W (2000) Prolegomenon to a theory of translation. The translation studies reader, 250-263
Gardner M, Grus J, Neumann M, Tafjord O, Dasigi P, Liu NF, Peters ME, Schmitz M, Zettlemoyer L (2018) Allennlp: A deep semantic natural language processing platform. Proceedings of Workshop for NLP Open Source Software (NLP-OSS)
Halliday M (1985) An introduction to functional grammar. Edward Arnold, London
Halliday MAK (1993) Towards a language-based theory of learning. Linguist. Educ. 5(2):93–116
Halliday MAK, Martin JR (1993) Writing science: Literacy and discursive power. pittsburgh press
Halliday MAK, Matthiessen C (2006) Construing experience through meaning: A language-based approach to cognition. Bloomsbury Publishing
Halverson SL (2003) The cognitive basis of translation universals. Target. Int. J. Transl Stud. 15(2):197–241
Halverson SL (2017) Gravitational pull in translation: Testing a revised model. Empirical translation studies: New methodological and theoretical traditions, 9-46
Han H, Jiang Y (2017) Rethinking translation in the light of complex adaptive system theory. Chin. Trans J. 38(02):19–24
Kenny D (2014) Lexis and creativity in translation: A corpus based approach. routledge
Kenny D (2017) Lexical hide-and-seek: Looking for creativity in a parallel corpus. In Intercultural faultlines (pp. 93-104). Routledge

Krüger R (2014) From s-explicitation to t-explicitation? Tracing the development of the explicitation concept. Across Lang. Cult. 15(2):153–175
Laviosa S (1998a) Core patterns of lexical use in a comparable corpus of english narrative prose. Meta 43(4):557–570
Laviosa S (1998b) The english comparable corpus: A resource and a methodology In Unity in diversity. Routledge
Laviosa S (2002) Corpus-based translation studies: Theory, findings, applications. Rodopi
Lin D (1998) An information-theoretic definition of similarity. ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning
Liu K, Ye R, Zhongzhu L, Ye R (2022) Entropy-based discrimination between translated chinese and original chinese using data mining techniques. Plos one 17(3):e0265633
Malmkjær K (1997) Punctuation in hans christian andersen's stories and in their translations into english. Benjamins Transl Libr. 17:151–162
Màrquez L, Carreras X, Litkowski KC, Stevenson S (2008) Semantic role labeling: An introduction to the special issue. Comput Linguist. 34(2):145–159
Martin JR (1992) English text: System and structure. John Benjamins
McEnery A, Xiao Z (2004) The lancaster corpus of mandarin chinese: A corpus for monolingual and contrastive language study. International Conference on Language Resources and Evaluation
McGrath D, Liardét C (2023) Grammatical metaphor across disciplines: Variation, frequency, and dispersion. Engl. Specif. Purp. 69:33–47
Miller GA (1995) Wordnet: A lexical database for english. Commun. ACM 38(11):39–41
Muñoz Martín R, Martín de León C (2020) Translation and cognitive science. In The routledge handbook of translation and cognition. Routledge
Olohan M (2003) How frequent are the contractions?: A study of contracted forms in the translational english corpus. Target. Int. J. Transl Stud. 15(1):59–89
Olohan M, Baker M (2000) Reporting that in translated english. Evidence for subconscious processes of explicitation? Across Lang. Cult. 1(2):141–158
Øverås L (1998) In search of the third code: An investigation of norms in literary translation. Meta 43(4):557–570
Pazienza MT, Pennacchiotti M, Zanzotto FM (2005) Textual entailment as syntactic graph distance: A rule based and a svm based approach. Proceedings of the First PASCAL Challenges Workshop on Recognizing Textual Entailment
Pradhan S, Ward W, Hacioglu K, Martin JH, Jurafsky D (2005) Semantic role labeling using different syntactic views. DBLP
Pym A (2005) Explaining explicitation. New trends in translation studies. In honour of Kinga Klaudy, 29-34
Reshmi SN, Shreelekshmi R (2019) Textual entailment based on semantic similarity using wordnet. 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)
Rong X (2014) Word2vec parameter learning explained. arXiv e-prints, arXiv: 1411.2738
Sang Z (2023) A neo-descriptivist approach to translation studies: Problems and methods (in chinese). J. Foreign Lang. 46(1):10
Shao Y, Liang C, Mao N (2012) The corpus construction and parsing technology based on chinese semantic dependency
Shi P, Lin J (2019) Simple Bert models for relation extraction and semantic role labeling. arXiv e-prints, arXiv: 1904.05255
Sweller J (2011) Cognitive load theory. In Mestre J P & Ross B H (Eds.), Psychology of learning and motivation. Elsevier Academic Press
Taverniers M (2006) Grammatical metaphor and lexical metaphor: Different perspectives on semantic variation. Neophilologus 90(2):321–332
Teich E (2003) Cross-linguistic variation in system and text: A methodology for the investigation of translations and comparable texts. Walter de Gruyter
Tirkkonen-Condit S (2004) Unique items — over- or under-represented in translated language?
Toury G (1995) Descriptive translation studies–and beyond. Benjamins Translation Library
Wu Z, Palmer M (1994) Verbs semantics and lexical selection. Proceedings of the 32nd annual meeting on Association for Computational Linguistics
Xiao R (2015) Source language interference in english-to-chinese translation. Yearbook of Corpus Linguistics and Pragmatics 2015: Current Approaches to Discourse and Translation Studies, 139-162
Xu X, Xu J (2021) Yiyan english-chinese parallel corpus (in chinese). Corpus Linguist. 1:3
Xue N, Palmer M (2009) Adding semantic roles to the chinese treebank. Nat. Lang. Eng. 15(1):143–172

Acknowledgements

The authors would like to thank Prof. Ruiying Yang and Ms. Haiyan Zhou for their inspiring advice and significant assistance during the revision process. This work was supported by the Humanities and Social Sciences Planning Fund of Ministry of Education, China (Grant No. 22YJAZH039).

Author contributions

Letao Wang: conceptualization and methodology, visualization, investigation, writing—original draft preparation, writing—reviewing and editing. Yue Jiang: writing—reviewing and editing, supervision.

Competing interests

The authors declare no competing interests.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Informed consents

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Correspondence and requests for materials should be addressed to Yue Jiang.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.