Behavioral/Cognitive

How Grammar Conveys Meaning: Language-Specific Spatial Encoding Patterns and Cross-Language Commonality in Higher-Order Neural Space

Jing Wang,[1,2,4] Hui Lin,[3] and Qing Cai[1,2,4,5]

[1]Key Laboratory of Brain Functional Genomics (MOE & STCSM), Affiliated Mental Health Center (ECNU), Institute of Brain and Education Innovation, School of Psychology and Cognitive Science, East China Normal University, Shanghai 20062, China, [2]Shanghai Changning Mental Health Center, Shanghai 200335, China, [3]Shanghai Key Laboratory of Artificial Intelligence in Learning and Cognitive Science, LAIX Inc., Shanghai 200090, China, [4]Shanghai Center for Brain Science and Brain-Inspired Technology, East China Normal University, Shanghai 200062, China, and [5]New York University-ECNU Institute of Brain and Cognitive Science, New York University, Shanghai 200062, China

Languages come in different forms but have shared meanings to convey. Some meanings are expressed by sentence structure and morphologic inflections rather than content words, such as indicating time frame using tense. This fMRI study investigates whether there is cross-language common representation of grammatical meanings that can be identified from neural signatures in the bilingual human brain. Based on the representations in intersentence neural similarity space, identifying grammatical construction of a sentence in one language by models trained on the other language resulted in reliable accuracy. By contrast, cross-language identification of grammatical construction by spatially matched activation patterns was only marginally accurate. Brain locations representing grammatical meaning in the two languages were interleaved in common regions bilaterally. The locations of voxels representing grammatical features in the second language were more varied across individuals than voxels representing the first language. These findings suggest grammatical meaning is represented by language-specific activation patterns, which is different from lexical semantics. Commonality of grammatical meaning is neurally reflected only in the interstimulus similarity space.

Key words: **fMRI; grammar; language; predictive model; semantics**

Significance Statement

Whether human brain encodes sentence-level meanings beyond content words in different languages similarly has been a long-standing question. We characterize the neural representations of similar grammatical meanings in different languages. Using complementary analytic approaches on fMRI data, we show that the same grammatical meaning is neurally represented as the common pattern of neural distances between sentences. The results suggest the possibility of identifying specific grammatical meaning expressed by different morphologic and syntactic implementations of different languages. The neural realization of grammatical meanings is constrained by the specific language being used, but the relationships between the neural representations of sentences are preserved across languages. These findings have some theoretical implications on a distinction between grammar and lexical meanings.

Received Mar. 31, 2023; revised Sep. 3, 2023; accepted Sep. 11, 2023.

Author contributions: J.W., H.L., and Q.C. designed research; J.W. performed research; J.W. contributed unpublished reagents/analytic tools; J.W. analyzed data; J.W. and Q.C. wrote the first draft of the paper; J.W. and Q.C. edited the paper; J.W. and Q.C. wrote the paper.

This work was supported by National Natural Science Foundation of China Grants 31970987 (to Q.C.) and 32100857 (to J.W.) and LAIX Inc. (Q.C.). We thank Dr. Samuel A. Nastase and two anonymous reviewers for their insightful comments and discussions that helped us improve the quality of this work. We also thank Miaomiao Zhu, Lechuan Wang, and Zhichao Wang for the assistance in data collection; Fan Yang and Guannan Zhao for recruiting participants and technical support; and Ruiqing Zhang for helpful discussion.

The authors declare no competing financial interests.

Correspondence should be addressed to Jing Wang at wangjing@psy.ecnu.edu.cn or Qing Cai at qcai@psy.ecnu.edu.cn.

https://doi.org/10.1523/JNEUROSCI.0599-23.2023

Introduction

Different languages share communicative intentions, while the linguistic forms expressing similar meanings may vary. Research over the past two decades has characterized how the human brain represents content words and their compositions in unprecedented detail (Mitchell et al., 2008; Bemis and Pylkkänen, 2011; Wang et al., 2017; Coutanche et al., 2020; Tang et al., 2023). Concepts of concrete objects and simple events are found represented by brain activation patterns that are common across different languages (Correia et al., 2014; Yang et al., 2017), indicating shared neural substrates for content-word-driven semantics.

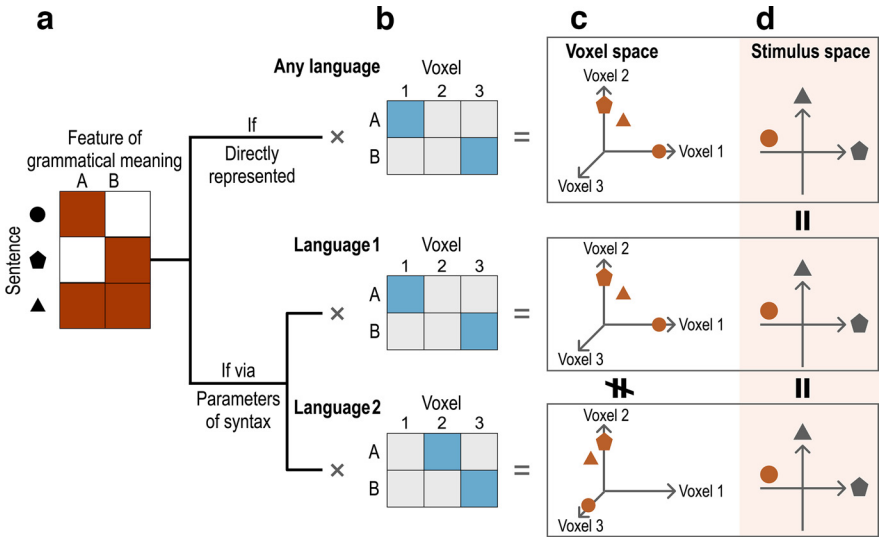

Figure 1. Conceptual illustration of common cross-language representations of grammatical meanings in neural similarity space regardless of commonality in the multivoxel activation patterns. ***a***, Assume three sentences, represented by the circle, the pentagon, and the triangle, vary on two grammatical meanings, A and B. The colors in the sentence-by-feature matrix indicate different values of the elements, namely, the extent to which a sentence expresses a grammatical meaning. While this study codes grammatical features binarily, the features can be continuous. ***b***, The grammatical meaning A or B may be neurally directly coded (upper branch) or via syntax (lower branch). In the second case (lower branch), two actual languages express the same meanings by different morphologic or syntactic properties, which are neurally realized by different sets of voxels, as illustrated by the blue squares. ***c***, Different hypotheses lead to different predictions on whether the same grammatical meaning is represented similarly in voxel space across language. ***d***, Nevertheless, the same underlying meaning representation in ***a*** can be captured when the neural representations are viewed in a space spanned by other sentences, i.e., the intersentence neural similarity. The coordinate systems represent the neural similarity space, in which the dimensions are stimuli and the location of a sentence indicates the neural similarities to other sentences.

On the other hand, certain meanings arise from morphosyntactic properties rather than content words (Goldberg, 1995). In sentences "I have walked my dog" and "I am walking my dog," different tenses and aspects indicate different time frames of the event. In Mandarin Chinese, time frame is expressed by using aspect markers such as "了" or adverbs such as "正在." Are these grammatical meanings, namely, meanings expressed by sentence structures, inflectional morphology, or function words, represented in common neural substrates across languages in a bilingual brain? While neural responses associated with different grammatical constructions have been identified (Allen et al., 2012), what remains unaddressed is whether such neural differences result from differences in formal linguistic structures, or from differences in grammatical meanings, such as the subtle semantic distinction between the ditransitive and the dative constructions (Pinker, 1991; Goldberg, 2003; Ambridge et al., 2012).

Because grammatical meanings are coupled with the specific morphologic and syntactic properties of a specific language, the present study uses different languages to identify the neural signature of meaning to dissociate it from formal syntactic properties. The question is whether the same grammatical meaning is represented by the same spatial neural activation pattern in different languages. One possibility is that particular grammatical meaning is encoded directly as particular neural activation pattern, just like lexical semantics is represented by concept-specific activation patterns regardless of the language (Buchweitz et al., 2012; Yang et al., 2017). If this is the case, it will be possible to identify grammatical meaning of one language based on the spatial patterns in another language.

A second possibility is that unlike lexical semantics, grammatical meaning is only a by-product of the representation of formal linguistic structure. In the view of generative grammar, the sets of parameters, or instantiations of the universal grammar, are different across languages, resulting in different syntactic structures (Chomsky, 1965, 1986). If generative mechanism is the only route to grammar representation (Marantz, 2005), we infer that the same grammatical meaning is not represented by the same neural activation patterns in two languages that do not share specific rules (using different "sets of parameters"). As long as the grammatical meanings are the same in two languages, one will observe between-language commonality in the second-order intersentence neural similarity space, where the intersentence similarity is the similarity of the spatial pattern of neural responses between pairs of the sentences within a language (Kriegeskorte et al., 2008; Haxby et al., 2020). Assume when a common grammatical meaning (Fig. 1*a*) is expressed in two different languages, the relevant grammatical properties in the two languages are neurally coded by different spatial patterns (Fig. 1*b*), resulting in different neural signatures of sentences with common grammatical meanings in different languages (Fig. 1*c*). Grammatical meaning cannot be decoded across languages by directly comparing the spatial patterns, but consistency will be observed when one compares the neural similarity patterns across sentences within a language with the similarity pattern of another language (Fig. 1*d*).

The present study investigates the neural representation of grammatical meaning by examining common grammatical meanings in different languages. We associate grammatical meaning with neural representation of sentences in neural similarity space or in voxel space, and test whether the learned mapping can be used to predict the neural signature of a sentence in one's second language given its grammatical meaning. This approach is applied to ensure the cross-language commonality can be accounted for by the hypothetical features of grammatical meanings, and to ensure that the learned mapping is generalizable to new sentences.

Materials and Methods

Participants

Forty healthy young adults were originally recruited from the East China Normal University community. One participant quit the experiment halfway, resulting in a total of 39 participants (25 females, age averaged

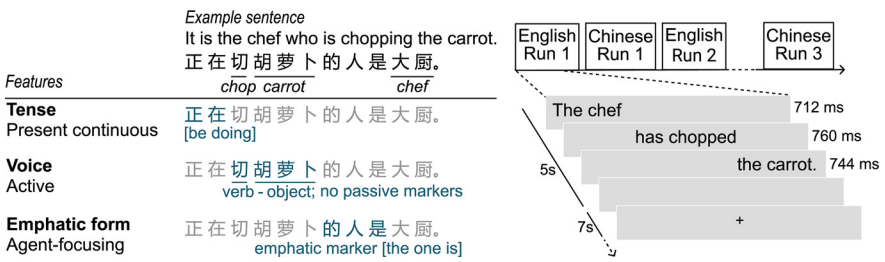

Figure 2. Example of stimulus sentence and schematic representation of the experimental paradigm. The left panel shows one sentence in English and its Chinese equivalent (Extended Data Fig. 2-1 for all the sentences). Colored texts mark the key components that indicated the each of grammatical features in that Chinese sentence. The right panel illustrates the experimental design. Participants read sentences silently during six runs of fMRI scans. Each sentence was presented one phrase at a time, with phrases cumulating from left to right on the screen. The presentation duration of each phrase was determined by the number of content words and number of letters.

at 21 years, SD = 2.54). Participants' age, gender, and language background were collected by questionnaire. Questions on participants' language experience were adapted from the Language History Questionnaire 2.0 (P. Li et al., 2014): "List the languages you have studied or learned, the age at which you started using each language in terms of listening, speaking, reading, and writing, and the total number of years you have spent using each language." English proficiency was indicated by the self-reported scores on the National College Entrance Examination of English. All the participants were Chinese-English late bilinguals. The mean age of onset for L2 was seven years (SD = 2.67). All the participants have been studying/using English for at least nine consecutive years. All the participants have passed the National College Entrance Examination of English. All participants provided written informed consent approved by the East China Normal University Institutional Review Board.

Experimental design and procedure

This study investigated fMRI-measured activation patterns when bilingual participants read sentences with nearly identical propositional contents but different grammatical constructions. Stimulus sentences consisted of 48 unique sentences: two languages (Chinese, English) × two tenses-aspects (present continuous, present perfect) × two voices (active, passive) × three emphatic forms (nonemphatic, agent-focusing, patient-focusing) × two scenarios (*Chef-chops-carrot*, *Grower-washes-tomato*). For example, "It is the carrot that the chef is chopping" is a patient-focusing sentence in English that describes one of the scenarios in present continuous tense-aspect and active voice (Extended Data Fig. 2-1). We chose these three grammatical areas for several reasons. First, these grammar features exist in both English and Chinese, i.e., both languages have these parameters "switched on." Second, these grammatical constructions are meaningful or interpretable outside the grammar system, as opposed to some other grammatical areas such as other subject-verb agreement. Third, as opposed to rules such as subject-verb-object that are common in both English and Chinese, the parameters of the selected rules, or morpho-syntactic structures, differ between the two languages, which ensures the perceptual dissimilarity between sentences in two languages (Fig. 2). Fourth, these three features are orthogonal to each other: manipulating one feature does not affect the validity of the others. Note that the active and passive voices depicted the same event (e.g., "The chef is chopping the carrot." vs "The carrot is being chopped by the chef.").

Below we used "grammatical feature" to refer to the features of tense, voice, and emphatic form, and used "grammatical construction" to refer to the specific combination of properties of these features in a specific sentence, such as "present continuous, active, and agent-focusing."

Before the MRI scans, participants read all the stimulus sentences and were informed that "These sentences described very similar events, but please try to comprehend the subtle differences in meanings." No participant reported issues regarding the sentences when they were asked whether the meaning and expression of any sentence were unclear or confusing. During the scan, participants silently read the sentences that were presented for a total of three times in six runs, three runs per language. Each run consisted of 24 sentence trials, a complete list of unique sentences in one language in a pseudorandomized order. In this slow event-related design, each sentence was presented one phrase at a time, with phrases cumulating from left to right on the screen (Fig. 2). Each phrase contained and only contained one content word (nouns or verbs), hence each sentence consisted of three phrases. The presentation duration of each phrase was computed as 600 ms × number of content words + 16 ms × number of letters for the English sentences. The presentation paradigm was motivated by the patterns of eye fixations during the reading of texts (Just and Carpenter, 1980). The basis timing of word display was modified from 300 to 600 ms based on preliminary tests on two late Chinese-English bilinguals to ensure sufficient time for processing. Because of the lack of corresponding research for the reading of Chinese, we applied the same equation when determining the timing of displaying Chinese sentences, i.e., 600 ms × number of content words + 16 ms × number of characters. The mean presentation duration of sentences was 2177 ms; SD = 208.35. The shortest presentation duration of a sentence was 1912 ms and the longest was 2552 ms. This range (640 ms) was shorter than 1/3 of the TR. After the last phrase of the sentence had been presented, a blank interval was presented to pad out the total duration of the sentence presentation to 5000 ms, during which the participants were allowed to finish their thoughts on the sentence. The blank interval has been found critical for semantic decoding and has been applied in previous fMRI studies (Wang et al., 2017; Yang et al., 2017). The blank interval was followed by a 7000-ms fixation cross ("+") presented in the center of the screen, during the presentation of which the participant was instructed to relax and clear their mind. Participants were instructed to think of the same properties of the person, object, and scenario described by a sentence each time the sentence was presented: "For example, when you encounter this sentence for the first time, if you happen to imagine the chef is cutting the carrot in half instead of cutting it into many slices, try thinking of that kind of carrot again – that is, a carrot in half – when you see the sentence for the second or the third time." Twelve additional trials were included, during which participants were asked to judge whether the present sentence used the same grammatical construction as the previous sentence. A 17-s fixation trial was presented at a random position in the stimulus sequence for each run, during which a "+" was shown at the center of the screen. Participants were instructed to relax and clear their mind while fixating on the sign.

fMRI imaging protocol

Functional and anatomic images were collected using a Siemens Prisma 3-T scanner with a 64-channel head coil at East China Normal University. Functional images were acquired using a gradient echo EPI pulse sequence with TR = 2200 ms, TE = 30 ms, and a 90° flip angle. Thirty-six 3.5-mm-thick AC-PC aligned slices were imaged with voxel size $3 \times 3 \times 3.5$ mm^3 FOV 192×192 mm^2. Structural images were acquired using a T1-weighted MPRAGE pulse sequence.

Image preprocessing

Preprocessing was performed using SPM12 (Wellcome Department of Cognitive Neurology, London, United Kingdom). Whole-cortex fMRI

data were corrected for slice timing and head motion. The structural image of each participant was coregistered to the mean of the functional image of the first scan session, tissue-segmented and bias-field corrected. The deformation information of normalizing the structural image to the Montreal Neurologic Institute (MNI) template was applied to spatially normalize the functional images. Functional images were then corrected for linear trend and low-frequency trends by applying a high-pass temporal filter at 0.0078 Hz. Further analyses were performed using in-house scripts on MATLAB7 (MathWorks).

For each presentation of a sentence, the percent signal change (PSC) was computed at each voxel in the brain image. The change was computed relative to a baseline activation level measured during and averaged over the 17-s fixation conditions. The baseline measurement started at 4 s after each fixation presentation onset to account for the hemodynamic response delay. The fMRI data of sentence reading consisted of the mean of three images, the first starting at 6.6 s from sentence onset. The PSC was then normalized to a mean of 0 and variance of 1 across sentences within each run (Pereira et al., 2009) to equate the overall intensities across scans and participants.

Control of features of no interests

Four types of feature sets that might correlate with the grammatical features were considered: presentation duration, visual-orthographic, phonological, and syntactic features. Presentation duration was the exact duration over all segments of each sentence. The first visual feature was orthographic complexity, measured as the number of strokes in a Chinese sentence or the number of letters in an English sentence, both being normalized within language. The second was number of words in a sentence. *Post hoc* examination showed no occipital or ventral temporal voxels were used in the grammatical coding (see "Distribution of informative voxels in two languages"), suggesting lower-level visual features were unlikely to confound the effect of interest.

Phonological features were taken into account despite the visual presentation paradigm because participants might generate covert speech during reading, and because phonological working memory demands might differ on sentences of different lengths. Twenty-five articulatory features were constructed following the rationale and approach of de Heer et al. (2017). Sentences were transcribed to phonemes. The occurrence of each articulatory feature in a sentence was coded according to the phoneme-articulation association defined by International Phonetic Association (https://www.internationalphoneticassociation.org/content/ipa-chart). These features represent both the speech sound characteristics and the vocal gesture characteristics (de Heer et al., 2017). Note that this feature set had coded the number of phonemes in a sentence, which was associated with the verbal working memory.

The first syntactic feature was syntactic complexity, measured by the node counts in the parse tree of a sentence (Miller and Chomsky, 1963; Frazier, 1985; Ferreira, 1991). Parsing was performed using the Stanford CoreNLP toolkit (Manning et al., 2014). The second syntactic feature was on word order, which binarily coded whether a sentence was agent-first or patient-first.

Thirty features of no interest were constructed. Because grammatical meanings were expressed by syntactic properties such as word order, function words, or inflections, which in turn affected other sentence properties the sentence length, and complexity, the features listed in this section were highly correlated with the grammatical meanings within language by nature. To control for the effects of confounders without eliminating the effects of interests, we only selected voxels that were better accounted for by grammatical features than features of no interest using the training data. A voxel was only selected for any following analysis if the grammatical feature set accounted for more variance (greater R^2) in that voxel than the all sets of features of no interest.

Voxel selection

Two-step voxel selection was performed using the training data before the model training. The first step was mandatory. As described above in Control of features of no interests, a voxel was retained only if the grammatical features could explain extra variance that were not explained by the features of no interests. *Post hoc* review showed that after the control of features of no interests without additional voxel selection, 2055–2535 voxels were modeled over cross-validation folds. The second step was optional and was referred to as "stability-based voxel selection." Data of three presentations of the same sentences in the training set were averaged to acquire a stable representation of the individual sentence. The sentence of the same grammatical attributes as the test sentence was ignored, resulting in 22 sentences. For example, if the test sentence was a present continuous, active, and agent-focusing sentence ("It is the chef who is chopping the carrot."), the other present continuous, active, and agent-focusing sentence that described the other scenario ("It is the grower who is washing the tomato.") was ignored. The 22 sentences formed 11 pairs, each pair being the two sentences of the same grammatical construction (e.g., present perfect, active, and agent-focusing) that described two different scenarios (one being *chef-chop-carrot* and the other being *grower-wash-tomato*). The grammatical tuning score of each voxel was computed as the Pearson correlation coefficient on the responses between the 11 pairs of sentences. This score indicated how well a voxel presented an activational profile over different grammatical constructions that were stable across different lexical contents. This voxel selection method was consistently applied to the following analyses. Results of using a range of different numbers of voxels were presented in the figures to show the robustness of the results. The numbers of voxels being selected were arbitrarily set to the rounded values of $30 \times 1.25^{n-1}$, where n was a natural number ranging from 1 to 20.

Predicting grammatical meaning signatures within language

We examined whether the responses in selected voxels to one grammatical construction was distinguishable from those to other constructions in the same language. A leave-one-sentence-out cross-validation procedure was applied to all the within-language analyses. In each cross-validation fold, the test image was the mean image over three presentation trials of the same sentence. The rest 69 trials of 23 sentences were used as the training data. A kernel ridge regression model (Hastie et al., 2009) was trained to learn the mapping from the grammar features to the response of each selected voxel. The independent variable were the 4-dimensional grammatical features. Each dimension binarily coded tense (present continuous vs present perfect), voice (active vs passive), whether it was an emphatic or nonemphatic sentence, or for an emphatic sentence, whether it was agent-focusing or patient-focusing. The dependent variable was the percent signal change at each selected voxel. The penalty weight of the ridge regression was chosen from a list of 23 candidate values that ranged from 10^{-7} to 10^{7}. The selection of parameters was performed using the generalized cross validation method for computational efficiency, which implicitly left one training sample out each time to train the model, applied it to the left-out sample, and calculated the prediction errors. The regularization parameter that provided the minimum mean squared error over all implicit folds was selected for each voxel.

The trained model was used to predict the neural signature associated with each of the 12 candidate grammatical constructions (two tenses × two voices × three emphatic forms), given its grammar feature coding. Only the voxels selected based on the training data were considered. The observed image of the left-out test sentence was compared with the 12 predicted images. The 12 candidates were ranked based on their cosine similarity scores to the actual test image. The performance of the model was assessed by the rank accuracy of correct grammatical construction, *(Number of candidates − Rank of the correct item)/(Number of candidates − 1)*.

For each of the prediction analyses, statistical significance of the resulting accuracy was determined based on the null distribution generated by a 10,000-iteration random permutation. The procedure in each iteration was the same as the main analysis, except that the labels of the entire data were randomly shuffled before further processing.

Identifying attributes of individual grammatical features

Performances on each of the three grammatical features were examined. For example, for tense prediction, the predicted images of all the present continuous sentences were averaged, and the predicted images of all present perfect sentences were averaged. The test image was compared with the two mean predicted images. Accuracy was determined by

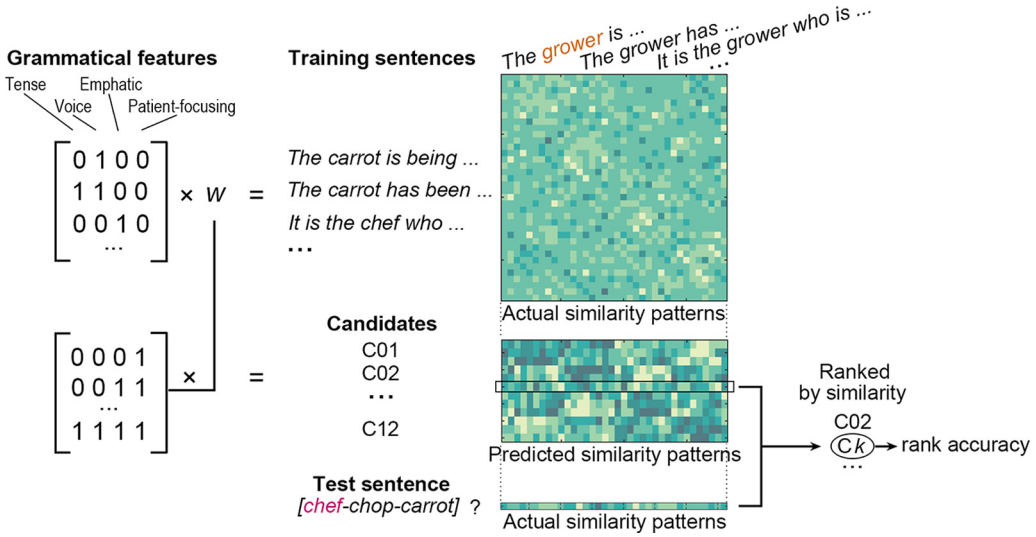

Figure 3. Illustration of methods of identifying grammatical construction in neural similarity space. When a *chef* sentence was left out as the test sentence, similarities of the activation patterns were computed between the other *chef* sentences and the *grower* sentences. The *chef* sentences served as the training exemplars and the *grower* sentences served as the neural feature space. Weights (*w*) that mapped the grammatical features to the neural similarities were learned and applied to generate predicted neural similarity patterns of 12 possible grammatical constructions. These predicted candidate patterns were compared with that of the test sentence. Normalized rank of the similarity of the actual construction over all candidates indicated the performance of the identification.

whether the predicted image with the correct tense label was closer to the test image.

Predicting grammatical meaning signatures across language on a voxel-by-voxel basis

In the voxel-wise cross-language prediction, the model was trained using PSCs associated with reading sentences in one language, to identify the grammatical construction of the PSC of each sentence in the other language. For this and the rest of the analyses, the procedures and algorithms applied were the same as in the within-language prediction unless specified. Training and testing data in the cross-language prediction were associated with two languages respectively.

Predicting grammar meaning signatures across language in neural similarity space

The model described below used the same training-testing protocol on the same sets of voxels as the voxel-based model described above. The neural signature of a sentence in this model was the intersentence similarity, namely, the Pearson's correlations of the sentence's percent signal change image to the images of all the sentences in the same language that described the other scenario. For instance, if the test sentence was about "chef chops carrot," the neural similarity was the correlations of the test sentence to each of the 12 sentences (including all three presentations) that described "grower washes tomato" (Fig. 3). The images contained only the voxels that were selected by the voxel selection procedure. The dependent variable of model training was each of these neural similarity scores. The independent variables were the 4-dimensional grammar features. The trained model was used to predict the neural similarity scores associated with each of the 12 possible grammatical constructions using the grammar features. The actual similarity vector of a test sentence was compared with the 12 predicted vectors. The performance of the prediction was evaluated by the rank accuracy of the correct grammatical construction over the 12 candidates. In short, the only difference between this similarity-based model and the voxel-based model described above was that the neural signature of a sentence, which was the voxel-wise responses in the voxel-based model, was now replaced with the neural similarity scores to other sentences.

Modeling neural similarity patterns using language-specific voxel selection

Voxel selection was performed independently for two languages. The selection procedure remained the same for the training data. For the test data, when one sentence was tested, the voxel selection was performed using all the other sentences of the same language as the test sentence and then applied to the test data. Note that the prediction was always performed independently on each test sentence, thus the data used for voxel selection were never involved as test items in this procedure.

Testing the role of features of no interests in explaining signals in selected voxels

This test was to further examine whether the select voxels were associated with the features of no interest. If the classification accuracy of grammatical meaning was contributed by the correlation between the confounders with grammatical meanings, these confounders should be able to predict the neural responses just like the grammatical feature vector did. This analysis applied the same procedure as the main analysis, except that the grammatical vector in the training and prediction was replaced with the 30-dimensional vector of features of no interest.

Language difference and individual differences in spatial patterns

Each participant had two binary maps of voxels being used in the modeling, one per language. The mean of cosine similarity between the maps of each pair of participants within a language was used as the estimate of intersubject similarity. Random permutation test was performed to compare the intersubject similarity between languages. In each iteration of a random permutation, the L1/L2 data label were randomly shuffled for each participant. If there was no systematic difference between languages (the null hypothesis), the intersubject similarity in the shuffled category would be statistically the same as the intersubject similarity within the real language category. The intersubject similarity in each shuffled category computed for 10,000 iterations formed the null-hypothesis distribution. The actual intersubject similarity within each language was test against the null distribution to obtain a *p* value.

Locating voxels that represented grammatical meanings in two languages

Post hoc examination showed when 437 voxels were selected in each participant in each language and included in the modeling, the mean accuracy across L1-to-L2 and L2-to-L1 directions was the highest among a range of numbers of voxels. Multikernel density analysis (Wager et al., 2009) was performed to identify voxels that were consistently selected across participants. Because the images were originally not smoothed for the purpose of retaining the spatial pattern of image, here the voxels were convolved with a 2-mm Gaussian kernel. For each language, a

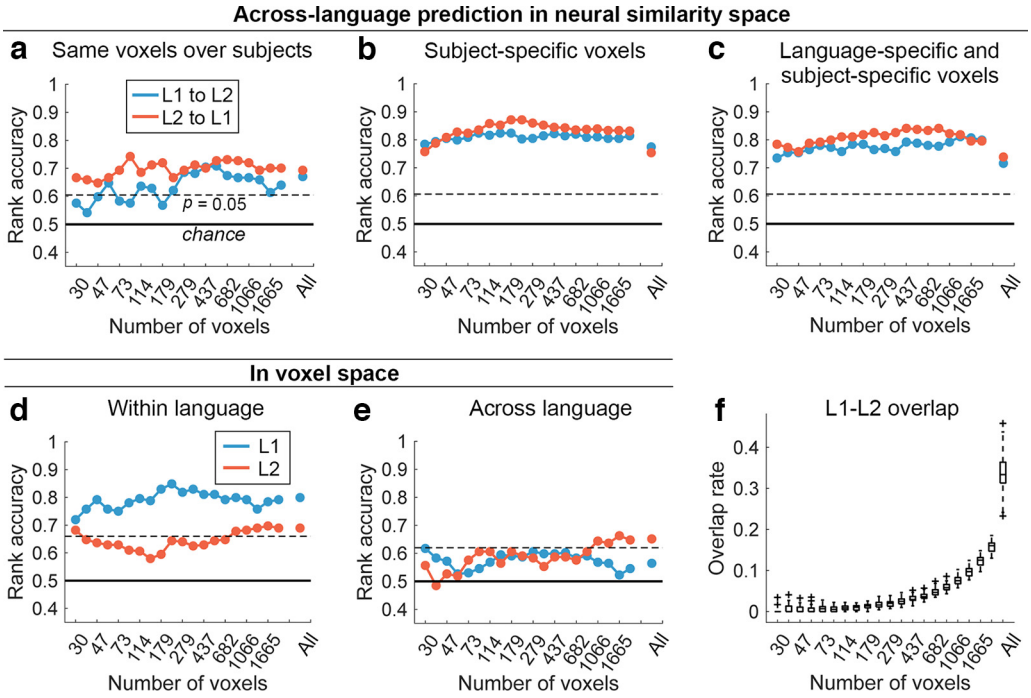

Figure 4. Results of using predicted neural signatures to identify grammatical construction of sentence. ***a***, Rank accuracy of cross-language identification of grammatical construction using neural similarity patterns (accuracies of identifying individual grammatical features were in Extended Data Fig. 4-1). Selected voxels were consistent across participants and languages. The *x*-axis indicated the number of voxels being used in the prediction. The "All" condition was when all the voxels that were better explained by the grammatical features than by the features of no interest were included. The dotted line indicated the most conservative critical value across models using different numbers of voxels ($p = 0.05$). ***b***, Rank accuracy of cross-language identification of grammatical construction using neural similarity patterns. The similarity patterns were computed using subject-specific voxels. For easier visualization of the comparison between Figure 4*a* and *b*, see Extended Data Figure 4-3. ***c***, Rank accuracy of cross-language identification of grammatical construction using neural similarity patterns. The similarity patterns were computed using subject-specific and language-specific voxels. ***d***, Rank accuracy of identifying grammatical construction within language in voxel space. ***e***, Rank accuracy of identifying grammatical construction across language in voxel space. At individual participant level, most participants with reliable within-language accuracy showed chance-level accuracy in voxel space and above-chance accuracy in similarity space for cross-language prediction (Extended Data Fig. 4-4). ***f***, Overlap rate of voxels being selected in two languages for identifying grammatical constructions. Each boxplot represents the distribution of all the participants. The overlapping voxels were sporadically located in multiple brain regions (Extended Data Fig. 4-2).

probability map was generated by averaging the convolved maps over participants. Null distributions were estimated by Monte Carlo simulation, in which the selected voxels were randomly located in the brain over 5000 iterations. The observed probability maps were then thresholded against the null distributions and corrected for multiple comparisons at p of 0.05.

Laterality differences in representing grammatical meanings in two languages

Laterality index (LI) of informative voxels within a language for each participant was computed as NL/(NL + NR), where NL was the number of voxels selected in the left hemisphere and NR was the number of voxels selected in the right hemisphere. LIs were computed at the whole brain level and within individual regions of interests using Automated Anatomical Labeling (Tzourio-Mazoyer et al., 2002). These regions have been consistently found responsive in various semantic tasks by meta-analysis (Binder et al., 2009) and were found consistently informative in the current study, namely, inferior frontal gyrus (pars opercularis, pars triangularis, and pars orbitalis), middle temporal gyrus/superior temporal sulcus, and the inferior parietal cortex (angular gyrus and supramarginal gyrus). For the within-region tests of LI differences between language, Bonferroni correction was applied to correct for the multiple comparisons.

Results

Predicting grammar-modulated neural signatures across language in voxel space

Voxel-wise cross-language prediction was performed to test whether the spatial activation patterns were similar between the sentences in different languages with common grammatical meanings. Activations of a voxel over sentences in one language were modeled by the vector of grammatical features. The trained model was tested by comparing the actual activation pattern of a sentence in the other language against the model-predicted signatures of all the possible candidate grammatical constructions. When all the voxels that were better explained by grammatical features than by all features of no interests were used, the mean accuracy of L1-to-L2 prediction was 0.56, not significantly higher than the chance-level accuracy ($p = 0.15$). The mean accuracy of L2-to-L1 prediction was 0.65 ($p = 0.009$). The significances of the accuracies were stably marginal when the different numbers of voxels were included, particularly for the L1-to-L2 prediction (Fig. 4*e*). Further investigation of the L2-to-L1 results suggested that the overall above-chance accuracy was likely to be mainly contributed by the emphatic form predictions: The accuracy of identifying emphatic versus nonemphatic sentences was 0.71 ($p = 0.0023$) and the accuracy of identifying agent-focusing versus patient-focusing sentences was 0.69 ($p = 0.06$; Extended Data Fig. 4-1*c*). These results suggested the representations in one language were not voxel-wise aligned to another, at least for the grammatical features of tense-aspects and voice.

Predicting grammar-modulated neural signatures across language in neural similarity space

We then examined the possibility of characterizing cross-language common representation of grammatical meanings based on intersentence neural similarities. As a secondary, higher-order

representation, interstimulus similarity can reveal the common representational pattern over different languages, regardless of whether the specific neural spatial patterns are the same or different across languages (Fig. 1). In this model, the neural signature for each sentence was a vector of its neural similarities to other sentences in the same language that used a different set of content words (Fig. 3).

The mean of the rank accuracy across sentences for cross-language prediction was reliably above chance level, being 0.67 ($p = 0.0025$, null-hypothesis distribution derived from 10,000 random permutations) when using model trained on first language (L1) to predict second language (L2) sentences (L1-to-L2), and being 0.69 ($p = 0.0003$) when using L2-based model to predict L1 sentences (L2-to-L1). Different models were also trained and tested on selected subsets of voxels. Additional stability-based voxel selection was performed in each cross-validation fold, which computed the signal correlation over pairs of sentences that were of the same attributes of tense, voice, and emphatic form but different content words (see Materials and Methods). The performance of L1-to-L2 prediction peaked at an accuracy of 0.71 ($p = 0.0001$) when using 546 voxels (Fig. 4*a*). The accuracy of L2-to-L1 prediction peaked at 0.74 ($p < 0.0001$) when using 92 voxels. Additional analyses on single grammatical features showed that tense, voice, or emphatic forms could be reliably identified at a wide range of number of voxels, in both L1-to-L2 and L2-to-L1 predictions (Extended Data Fig. 4-1*b*), which suggested the identifiability of the entire-sentence grammatical construction (Fig. 4) was not driven by a single grammatical feature.

Therefore, the modeling of neural similarities is more generalizable to a different language compared with the voxel-wise modeling. Cross-language commonalities in representing grammatical meanings are shown in intersentence neural similarity relations. The voxel activation pattern associated with an English sentence in passive voice and present continuous tense may differ from its counterpart in Chinese, but the position of the sentence in a space spanned by the neural signatures of other English sentences is analogous to the relative neural distances of its Chinese equivalent against other Chinese sentences.

Predicting grammar-modulated neural signatures within one language in voxel space

To verify that grammatical features were represented by voxel responses in a language, the within-language prediction was performed using a cross-validation protocol, in which the data of each sentence was left out at a time for validation and the rest of the sentences in the same language were used to train the model. When all the voxels that were better explained by grammatical features than by features of no interests were included, the average rank accuracy of identifying the correct attributes of a test sentence was reliably above chance in each language, being 0.80 ($p = 0.0009$) in L1 and 0.69 in L2 ($p = 0.0296$; Fig. 4*d*). Direct comparison on the mean of the accuracies of within-language and cross-language prediction in voxel space showed that the within-language accuracy was significantly higher than the cross-language accuracy, Wilcoxon rank sum test, rank sum = 687.5, $p = 0.02$. Hence, although all the sentences depicted highly similar scenarios using the same sets of content words, the sentence identities were still reliably classified for either language. By using the same set of grammatical features, it was possible to explain both the neural signatures of reading Chinese and reading English sentences.

Spatial variability across individuals and languages

Results by far suggested that grammatical features within a language were identifiable in voxel space, but the representations of grammatical meanings were not voxel-wise aligned between languages. We first tested whether the mismatch of voxels occurred at nearby locations by examining the association between the spatial proximity of voxels and the encoding patterns between languages, following the approach of Guntupalli et al. (2016). The weight similarity between languages was computed for a given pair of brain locations (voxels) using Pearson's correlation within each participant. Pearson's correlation between the weight similarity and the Euclidean distance of voxel pair was then computed to examine whether greater distance was associated with greater weight dissimilarity. For all pairs of voxels, the mean distance-weight correlation was -9×10^{-4} over participants, ranging from −0.046 to 0.023. When only the voxel pairs that were 0 (same brain location for two languages) to four voxels away were considered, the weight similarity did not decrease when voxel distance increased: the mean correlation over participants was 7×10^{-6}, ranging from −0.068 to 0.085 (Extended Data Fig. 5-1). The low correlation implied that voxels with matching weights were not at nearby locations. We then examined whether allowing for the voxel-wise misalignment across participants or across language when constructing neural similarity patterns would also result in reliable cross-language identifications.

Taking individual differences into account

To take into account the individual differences in either structure or functional topographies, the intersentence neural similarity was computed within individual participants and then averaged across participants. The rank accuracy was 0.77 ($p < 0.0001$) for L1-to-L2 prediction and 0.75 ($p < 0.0001$) for L2-to-L1 prediction. When different numbers of voxels were searched though, the performance of L1-to-L2 prediction peaked at 0.82 ($p < 0.0001$) and the accuracy of L2-to-L1 prediction peaked at 0.87 ($p < 0.0001$; Fig. 4*b*; Extended Data Fig. 4-3).

Taking language differences into account

Optimizing voxel selection within each language separately resulted in similar performances in characterizing the common grammatical representations between languages. Here, we constructed the similarity vectors for a test sentence based on the voxels selected using other sentences in the same language as the test sentence. The resulting accuracies ranged from 0.72 to 0.81 in L1-to-L2 prediction and ranged from 0.67 to 0.75 when training on L2 to predict L1 (Fig. 4*c*). *Post hoc* examination showed that the cross-language commonality of neural representations could be captured with a small amount of overlap of voxels (Fig. 4*f*; Extended Data Fig. 4-2). When the mean cross-language prediction accuracy across the two directions was maximal, being 0.86 ($p < 0.0001$) and 0.88 ($p < 0.0001$) in either direction, the overlap rates of voxels were 0.01%, i.e., 0.01% of voxels were selected by both languages over all the selected voxels.

Language difference and individual differences in spatial patterns

We then asked whether the between-language difference in spatial pattern was systematic given the intersubject variability within language. The intersubject similarity of the binary maps of voxel selection within each language was tested against the empirically generated distribution of intersubject similarity when the language labels (L1 or L2) for randomly selected maps were switched (see Materials and Methods). intersubject cosine similarity within L1 was 0.49, significantly larger than

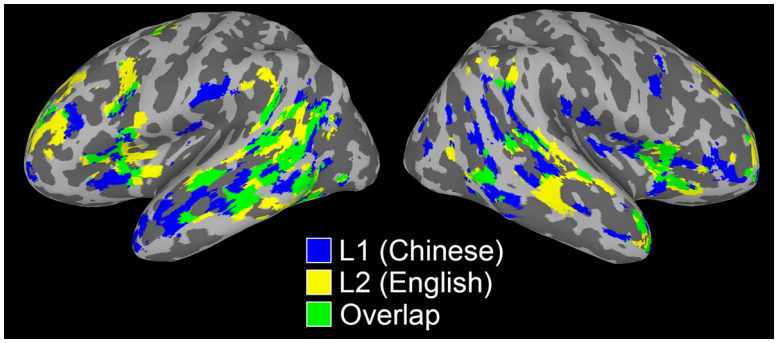

Figure 5. Spatially interleaved distribution of informative voxels for identifying grammatical meanings in each language. The colored voxels were consistently (simulation-based $p = 0.05$) selected across participants for modeling grammatical features. The association of the voxels' distance with the weight similarity was negligible (Extended Data Fig. 5-1). Voxels representing Chinese (L1) sentences were more right-lateralized than those over English (L2) sentences (Extended Data Fig. 5-2). Voxels representing L2 grammatical features were more varied across participants (Extended Data Figs. 5-3, 5-4).

the random-shuffling distribution ($p = 0.0063$). By contrast, the intersubject similarity in L2 was 0.45, the p value of which against the random-shuffling distribution was 0.9992, suggesting the within-L2 similarity was smaller than L1. Comparison of the intersubject similarity within L1 versus the similarity within L2 showed a significant difference between languages (paired-sample $t_{(740)} = 16.24$, $p = 5.98 \times 10^{-51}$; Extended Data Fig. 5-3). Therefore, the locations of voxels representing L2 grammatical features were more varied across individuals than the locations of voxels representing L1.

Between-language comparison of the intersubject similarity was further performed within individual brain regions defined by the Automated Anatomical Labeling (Tzourio-Mazoyer et al., 2002). Only the 90 noncerebellum regions were considered. Forty-four of the 90 regions showed significant differences between languages (FWE corrected $p = 0.05$), including bilateral frontal and temporal cortices, supramarginal gyri, anterior cingulate cortex, and precuneus (Extended Data Fig. 5-4). Thus, widespread brain regions contributed to the greater intersubject variability of L2 representation identified at the whole-brain level.

Distribution of informative voxels in two languages

The nonoverlapping voxels in two languages might be located in discrete brain regions, or located in common areas in an interwoven way, presenting different local spatial patterns. If the latter was true, the informative brain areas in the two languages were expected to overlap after slight smoothing. To characterize how the neural representations of grammatical meanings differ between languages, multikernel density analysis (Wager et al., 2009) was performed to identify consistently selected voxels across participants ($p = 0.05$, FWE corrected). The resulting maps of brain regions for the two languages revealed an interleaved pattern (Fig. 5). Overlapped brain areas across languages were identified in bilateral superior temporal sulcus, posterior inferior temporal gyrus, posterior supramarginal gyrus, insula, inferior frontal gyrus, middle frontal gyrus, middle cingulate cortex, and anterior cingulate cortex. The results suggested that the voxels selected for representing grammatical meanings in the two languages occupied common neural system but different local patterns.

Laterality differences in representing grammatical meanings in two languages

Informative voxels that represented grammatical features were bilaterally located for both Chinese and English sentences (Fig. 5). When the number of selected informative voxels was compared between hemispheres at the whole brain level, a small but robust lateralization difference was found at multiple thresholds of the numbers of selected voxels: voxels with stable response profiles over Chinese (L1) sentences were more right-lateralized compared with those over English (L2) sentences (Extended Data Fig. 5-2*a*). Laterality was then examined within each of the following regions that have been identified by meta-analysis (Binder et al., 2009) and were found consistently informative in the current study: inferior frontal gyrus, superior temporal sulcus/middle temporal gyrus, and the inferior parietal cortices including angular gyrus and supramarginal gyrus. We particularly tested the between-languages difference when 437 voxels were selected in each participant in each language, because this number of selected voxels results in the highest cross-language accuracy among the range of numbers of voxels (Fig. 4). Only the inferior parietal lobule showed a significant between-languages difference in laterality index, but it was the L1 that showed stronger left lateralization ($t = 4.00$, FWE corrected $p = 0.0018$). The pattern was consistently seen when multiple numbers of selected voxels were examined (Extended Data Fig. 5-2*b*).

Identifying grammatical constructions at individual participant level

The prediction of grammar-modulated neural signatures was performed within individual participants using the same approach as when using all participants' data. The grammatical constructions were reliably classified in 27 out of 39 participants for L1 sentences, and were reliably classified in 26 participants for L2 sentences (Extended Data Fig. 4-4*a*). Sixteen participants showed significant accuracy in both languages. These single-subject-level performances were comparable to those in previous fMRI decoding studies (see Discussion for details). Among these 16 participants, the accuracy of cross-language prediction in similarity space was significantly above chance level in 13 participants when models were trained on L1 to predict L2, and in 14 participants when trained on L2 to predict L1 sentences. By contrast, among the 16 participants, the accuracy of cross-language prediction in voxel space was significantly above chance level in only one participant when models were trained on L1 to predict L2, and in 3 participants when trained on L2 to predict L1 sentences (Extended Data Fig. 4-4*b*). For the three participants who did not show reliable accuracy when the prediction was performed in similarity space in either direction of prediction, none showed reliable accuracy in voxel space. The mean accuracy of

the prediction in similarity space was significantly higher than that in voxel space (Wilcoxon signed-rank test, two-sided $p = 0.0004$ for training on L1 to predict L2; two-sided $p = 0.0004$ for training on L2 to predict L1). These results were consistent with the group-level findings that cross-language common representation of grammatical meanings was shown in neural similarity relations but not in voxel-to-voxel relations.

Testing the role of features of no interests in explaining signals in selected voxels

To ensure that the voxels identified for representing grammatical features were not explained the features of no interest, the sentence identification in the neural similarity space was performed by training models using the 30-dimensional features of no interests. The resulting rank accuracy was 0.45 for L1-to-L2 prediction and 0.36 for L2-to-L1 prediction, none being significantly greater than chance level ($ps > 0.05$), suggesting that these voxels did not use the potential confounding variables to achieve cross-language sentence prediction.

Discussion

This study investigated how meanings conveyed by morphologic and syntactic features of different languages were represented in the bilingual brain. On a coarse scale, brain regions that were consistently responsive to different morphosyntactic features in one language were spatially interleaved with those in another language. On a finer scale, the spatial patterns that represented specific grammatical meanings differed between languages, as suggested by the marginal accuracy in cross-language meaning identification. Cross-language decoding based on neural similarities resulted in robust above-chance accuracy, suggesting common pattern of intersentence relations across languages. Substantial locational overlap was not necessary for capturing the cross-language representational commonalities. Thus, common grammatical meaning in different languages is represented by distinct neural spatial patterns, and is aligned through higher-order neural similarity space (Fig. 1).

Regarding how the two languages differed in the representational patterns, first, locations of voxels that represented sentence constructions in L2 showed greater variability across individuals than L1. Considerable individual differences has been found in the distributed native language network (Fedorenko et al., 2010; Braga et al., 2020). The present result suggests greater cross-individual variability for representing grammatical constructions in second language in a large number of brain regions. Second, despite the intersubject differences, cross-participant commonalities were still identified in regions that composed the universal language network (Malik-Moraleda et al., 2022). The overall distribution of the voxels in two languages lay in similar brain regions, so the spatial pattern differences were local rather than regional. Third, although at the whole-brain level, informative voxels were right-lateralized for representing Chinese and left-lateralized for representing English, greater left lateralization for Chinese was found within the semantic systems, specifically in the inferior parietal lobule. Previous studies have reached little consensus on whether and how the second language representation is different from L1 representation (Liu and Cao, 2016; Cargnelutti et al., 2019; Sulpizio et al., 2020; H. Li et al., 2021). This study suggests that for representing grammatical meanings, it might be the discrepancy in L2 representation across individuals that resulted in the between-language discrepancy at group level.

The present evidence implies that the spatial patterns of neural activations for representing grammatical meaning is language-specific, which is different from content-word semantics. Regardless of the debate on whether grammar is an autonomous linguistic structure, the morphologic inflections and sentence structures do bear semantic information. Formal linguistics and cognitive grammar theory hold different views on whether grammar forms an independent level of representation distinct from semantics. Previous studies have identified common neural signatures of concepts expressed by content words and their simple compositions (Buchweitz et al., 2012; Correia et al., 2014; Zinszer et al., 2016; Yang et al., 2017). This study shows a different pattern for representing grammatical meanings, specifically, different languages realize grammatical meanings by different voxel activation patterns. Such differences might be derived from the differences in sentence structures or in the linguistic units in the two languages, namely, the use of morphologic inflections or function words. Moreover, the fact that the grammatical meanings in two languages were aligned in the secondary space of neural similarity indicated structures did contain meanings. Such differential neural signatures captured by the models are likely to represent the grammatical meanings, rather than the structural complexity of sentences, for several reasons. First, the grammatical features were defined and aligned between languages by meanings rather than surface structures, in that the forms or structures in the two languages were different when representing common grammatical meanings. For instance, the word orders of the same grammatical meaning in the two languages are different in several cases: patient-focusing sentences in English in this study use the inverted structure that moves the object forward, whereas patient-focusing sentences in Chinese keep the object at the end of the sentences. Hence, the cross-language sentence prediction is a reliable test of the existence of neural representation of grammar-determined sentence meanings rather than structures. Second, the effects of various features of no interests were controlled throughout the analyses and explicitly tested. Third, the stimulus sentences were simple by design and expected to elicit no comprehension difficulty. Fourth, participants were proficient late bilinguals; they were familiarized with the sentences before they went into the scanner and reported no questions or doubts when asked. In addition, the distinct locations that were responsive to the differences in tense, voice, or emphatic form may reflect the distributed representations of semantics determined by different grammar features, rather than a single variable such as sentence complexity.

Performances of the within-language prediction were above chance-level on about two-thirds (67%) of the participants. Because above-chance performance for single participant's model is not a necessity for making inferences based on a reasonable sample size, the majority of the studies did not report the subject-level results (Baron and Osherson, 2011; Coutanche and Thompson-Schill, 2015; Parkinson et al., 2014; Yang et al., 2017; Vodrahalli et al., 2018; Weisberg et al., 2018; Frankland and Greene, 2020). For studies that reported subject-level semantic decoding performances, including word-level decoding, none to our knowledge used >20 participants a single decoding task, whereas the present study included 39 participants. Many of the studies aimed to classify word-level concepts or semantic categories (Mitchell et al., 2008; Mason and Just, 2016; Bauer and Just, 2017; Vargas and Just, 2022). Among the studies that decoded sentence-level semantics at individual level, one had a sample size of 20 participants, thirteen of whom (65%) yielded above-chance accuracy in either a sentence classification or a word

classification task (Allen et al., 2012). Other two studies included small samples (six to eight subjects in a task), acquired large within-subject dataset (over 4 h of scan per subject), and yielded above-chance accuracy on all subjects (Wang et al., 2017; Pereira et al., 2018). Note in both studies, the participants were specifically recruited based on the decoding results of their data in other tasks. One of the studies (Pereira et al., 2018) also reported word-level decoding results on 16 participants, where 7 out of 16 (44%) or 10 out of 16 (63%) participants showed significant accuracy depending on the task difficulty. We contend that it is typical that the group-level result is not replicated on every single subject, and it is unnecessary to replicate the group-level results on each subject to make inferences when the sample size is sufficient.

Some limitations of this study were related to the use of unnaturalistic language stimuli with very limited samples: we only investigated a small number of grammatical meanings within a fixed word-semantic space (chef cutting carrot/grower washing tomatoes). Because a specific grammatical meaning is realized by a specific sentence construction in this study, the frequency of sentence construction might be a potential confounder to the representation of grammatical meanings. Future studies are required to test whether the current findings apply in rich semantic contexts and in naturalistic language processing.

The power of human language in expressing unlimited number of thoughts comes from the flexible manipulations of a limited set of linguistic constituents, the complexity of which goes beyond superimpositions of multiword semantics. This study showed the possibility of systematically examining brain representations of the conceptualization of grammars. What we have discovered in this study suggests separable neural response patterns for the representation of sentence structures for two languages, yet the between-language differences are systematically driven by meaning, which brings the neural representations to convergence in the higher-order space.

References

Allen K, Pereira F, Botvinick M, Goldberg AE (2012) Distinguishing grammatical constructions with fMRI pattern analysis. Brain Lang 123: 174–182.

Ambridge B, Pine JM, Rowland CF, Chang F (2012) The roles of verb semantics, entrenchment, and morphophonology in the retreat from dative argument-structure overgeneralization errors. Language 88:45–81.

Baron SG, Osherson D (2011) Evidence for conceptual combination in the left anterior temporal lobe. Neuroimage 55:1847–1852.

Bauer AJ, Just MA (2017) A brain-based account of "basic-level" concepts. Neuroimage 161:196–205.

Bemis DK, Pylkkänen L (2011) Simple composition: a magnetoencephalography investigation into the comprehension of minimal linguistic phrases. J Neurosci 31:2801–2814.

Binder JR, Desai RH, Graves WW, Conant LL (2009) Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cereb Cortex 19:2767–2796.

Braga RM, DiNicola LM, Becker HC, Buckner RL (2020) Situating the left-lateralized language network in the broader organization of multiple specialized large-scale distributed networks. J Neurophysiol 124:1415–1448.

Buchweitz A, Shinkareva SV, Mason RA, Mitchell TM, Just MA (2012) Identifying bilingual semantic neural representations across languages. Brain Lang 120:282–289.

Cargnelutti E, Tomasino B, Fabbro F (2019) Language brain representation in bilinguals with different age of appropriation and proficiency of the second language: a meta-analysis of functional imaging studies. Front Hum Neurosci 13:154.

Chomsky N (1965) Aspects of the theory of syntax. Cambridge: The MIT Press.

Chomsky N (1986) Knowledge of language: its nature, origin, and use. Westport: Praeger.

Correia J, Formisano E, Valente G, Hausfeld L, Jansma B, Bonte M (2014) Brain-based translation: fMRI decoding of spoken words in bilinguals reveals language-independent semantic representations in anterior temporal lobe. J Neurosci 34:332–338.

Coutanche MN, Thompson-Schill SL (2015) Creating concepts from converging features in human cortex. Cereb Cortex 25:2584–2593.

Coutanche MN, Solomon SH, Thompson-Schill SL (2020) Conceptual combination. In: The cognitive neurosciences, Ed 6 (Poeppel D, Mangun GR, and Gazzaniga MS, eds). Cambridge: The MIT Press.

de Heer WA, Huth AG, Griffiths TL, Gallant JL, Theunissen FE, Heer WA, de Huth AG, Griffiths TL, Gallant JL, Theunissen FE (2017) The hierarchical cortical organization of human speech processing. J Neurosci 37:6539–6557.

Fedorenko E, Hsieh P-J, Nieto-Castañón A, Whitfield-Gabrieli S, Kanwisher N (2010) New method for fMRI investigations of language: defining ROIs functionally in individual subjects. J Neurophysiol 104:1177–1194.

Ferreira F (1991) Effects of length and syntactic complexity on initiation times for prepared utterances. J Mem Lang 30:210–233.

Frankland SM, Greene JD (2020) Two ways to build a thought: distinct forms of compositional semantic representation across brain regions. Cereb Cortex 30:3838–3855.

Frazier L (1985) Syntactic complexity. In: Natural language parsing: psychological, computational, and theoretical perspectives (Dowty DR, Karttunen L, and Zwicky AM, eds), pp 129–189. Cambridge: Cambridge University Press.

Goldberg AE (1995) Constructions: a construction grammar approach to argument structure. Chicago: Chicago University Press.

Goldberg AE (2003) Constructions: a new theoretical approach to language. Trends Cogn Sci 7:219–224.

Guntupalli JS, Hanke M, Halchenko YO, Connolly AC, Ramadge PJ, Haxby JV (2016) A model of representational spaces in human cortex. Cereb Cortex 26:2919–2934.

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction, Ed 2. San Diego: Springer.

Haxby JV, Guntupalli JS, Nastase SA, Feilong M (2020) Hyperalignment: modeling shared information encoded in idiosyncratic cortical topographies. Elife 9:e56601.

Just MA, Carpenter PA (1980) A theory of reading: from eye fixations to comprehension. Psychol Rev 87:329–354.

Kriegeskorte N, Mur M, Bandettini P (2008) Representational similarity analysis – connecting the branches of systems neuroscience. Front Syst Neurosci 2:4

Li H, Zhang J, Ding G (2021) Reading across writing systems: a meta-analysis of the neural correlates for first and second language reading. Bilingualism 24:537–548.

Li P, Zhang F, Tsai E, Puls B (2014) Language history questionnaire (LHQ 2.0): a new dynamic web-based research tool. Bilingualism 17:673–680.

Liu H, Cao F (2016) L1 and L2 processing in the bilingual brain: a meta-analysis of neuroimaging studies. Brain Lang 159:60–73.

Malik-Moraleda S, Ayyash D, Gallée J, Affourtit J, Hoffmann M, Mineroff Z, Jouravlev O, Fedorenko E (2022) An investigation across 45 languages and 12 language families reveals a universal language network. Nat Neurosci 25:1014–1019.

Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D (2014) The Stanford CoreNLP natural language processing toolkit. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp 55–60. Baltimore, Maryland.

Marantz A (2005) Generative linguistics within the cognitive neuroscience of language. Linguist Rev 22:429–445.

Mason RA, Just MA (2016) Neural representations of physics concepts. Psychol Sci 27:904–913.

Miller G, Chomsky N (1963) Finitary models of language users. In: Handbook of mathematical psychology (Luce RD, Bush RR, and Galanter E, eds). New York: Wiley.

Mitchell TM, Shinkareva SV, Carlson A, Chang K-M, Malave VL, Mason RA, Just MA (2008) Predicting human brain activity associated with the meanings of nouns. Science 320:1191–1195.

Parkinson C, Liu S, Wheatley T (2014) A common cortical metric for spatial, temporal, and social distance. J Neurosci 34:1979–1987.

Pereira F, Mitchell T, Botvinick M (2009) Machine learning classifiers and fMRI: a tutorial overview. Neuroimage 45:S199–S209.

Pereira F, Lou B, Pritchett B, Ritter S, Gershman SJ, Kanwisher N, Botvinick M, Fedorenko E (2018) Toward a universal decoder of linguistic meaning from brain activation. Nat Commun 9:963.

Pinker S (1991) Learnability and cognition: the acquisition of argument structure. Cambridge: The MIT Press.

Sulpizio S, Del Maschio N, Fedeli D, Abutalebi J (2020) Bilingual language processing: a meta-analysis of functional neuroimaging studies. Neurosci Biobehav Rev 108:834–853.

Tang J, LeBel A, Jain S, Huth AG (2023) Semantic reconstruction of continuous language from non-invasive brain recordings. Nat Neurosci 26:858–866.

Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, Mazoyer B, Joliot M (2002) Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage 15:273–289.

Vargas R, Just MA (2022) Similarities and differences in the neural representations of abstract concepts across English and Mandarin. Hum Brain Mapp 43:3195–3206.

Vodrahalli K, Chen P-H, Liang Y, Baldassano C, Chen J, Yong E, Honey C, Hasson U, Ramadge P, Norman KA, Arora S (2018) Mapping between fMRI responses to movies and their natural language annotations. Neuroimage 180:223–231.

Wager TD, Lindquist MA, Nichols TE, Kober H, Van Snellenberg JX (2009) Evaluating the consistency and specificity of neuroimaging data using meta-analysis. Neuroimage 45:S210–S221.

Wang J, Cherkassky VL, Just MA (2017) Predicting the brain activation pattern associated with the propositional content of a sentence: modeling neural representations of events and states. Hum Brain Mapp 38:4865–4881.

Weisberg SM, Marchette SA, Chatterjee A (2018) Behavioral and neural representations of spatial directions across words, schemas, and images. J Neurosci 38:4996– 5007.

Yang Y, Wang J, Bailer C, Cherkassky V, Just MA (2017) Commonality of neural representations of sentences across languages: predicting brain activation during Portuguese sentence comprehension using an English-based model of brain function. Neuroimage 146:658–666.

Zinszer BD, Anderson AJ, Kang O, Wheatley T, Raizada RDS (2016) Semantic structural alignment of neural representational spaces enables translation between English and Chinese words. J Cogn Neurosci 28:1749–1759.