CAMBRIDGE
UNIVERSITY PRESS

ARTICLE

Linguistic synesthesia detection: Leveraging culturally enriched linguistic features

Qingqing Zhao[1],[†] (ID), Yunfei Long[2],[†] (ID), Xiaotong Jiang[3], Zhongqing Wang[3], Chu-Ren Huang[4] (ID) and Guodong Zhou[5]

[1]Institute of Linguistics, Chinese Academy of Social Sciences, Beijing, China, [2]School of Computer Science and Electronic Engineering, University of Essex, Essex, UK, [3]Natural Language Processing Lab, Soochow University, Suzhou, China, [4]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong, China, and [5]School of Computer Science and Technology, Soochow University, Suzhou, China

Corresponding authors: Qingqing Zhao; Email: zhaoqq@cass.org.cn; Chu-Ren Huang; Email: churen.huang@polyu.edu.hk

Abstract

Linguistic synesthesia as a productive figurative language usage has received little attention in the field of Natural Language Processing (NLP). Although linguistic synesthesia is similar to metaphor concerning involving conceptual mappings and showing great usefulness in the NLP tasks such as sentiment analysis and stance detection, the well-studied methods of metaphor detection cannot be applied to the detection of linguistic synesthesia directly. This study incorporates comprehensive linguistic features (i.e., character and radical information, word segmentation information, and part-of-speech tagging) into a neural model to detect linguistic synesthetic usages in a sentence automatically. In particular, we employ a span-based boundary detection model to extract sensory words. In addition, a joint model is proposed to detect the original and synesthetic modalities of the sensory words collectively. Based on the experiments, our model is shown to achieve state-of-the-art results on the dataset for linguistic synesthesia detection. The results prove that leveraging culturally enriched linguistic features and joint learning are effective in linguistic synesthesia detection. Furthermore, as the proposed model leverages non-language-specific linguistic features, the model would be applied to the detection of linguistic synesthesia in other languages.

Keywords: linguistic synesthesia; linguistic features; a neural network model; Chinese

1. Introduction

Processing of figurative languages has been one of the most challenging tasks in Natural Language Processing (NLP). Among the different types of figurative meanings, metaphor and irony have been studied extensively in NLP. The processing of metaphor and irony has been shown to make significant contributions to tasks such as semantic parsing, sentiment and emotion analysis, stance detection, etc. (Weitzel, Prati, and Aguiar 2016; Hercig and Lenc 2017; Zhang *et al.* 2019; Su, Wu, and Chen 2021). However, linguistic synesthesia, as one of the most productive and frequently used figurative languages, has not received much attention in computational linguistics so far.

Linguistic synesthesia is the use of words and expressions from one sensory modality to describe concepts in a different sensory modality (Ullmann 1957; Williams 1976; Shen 1997). Examples below illustrate the usages of linguistic synesthesia in English, Mandarin, and Turkish respectively.

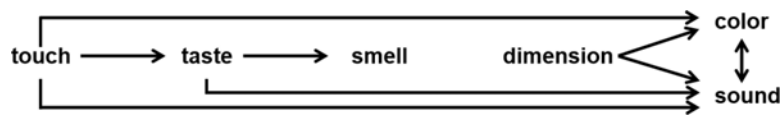[†]These authors contributed equally to this work.

Figure 1. A hierarchical model for linguistic synesthesia (Williams 1976, see p. 463).
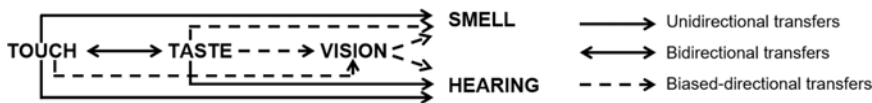

Figure 2. Transfer directionalities of linguistic synesthesia based on Mandarin corpus data (Zhao *et al.* 2019, see p. 9).

- Analogue recordings sound[HEARING/target] warmer[TOUCH/source] than digital. (Strik Lievers 2015, see p. 79)

- 葉色 *ye4se4*   濃 *nong2*[TASTE/source]   綠 *lv4*[VISION/target]
  leaf-color   of intense taste   green
  "The color of leaves is deep green."
  (Zhao, Huang, and Long 2018, see pp. 1178-1179)

- yagmur-un   ıslak[TOUCH/source]   koku-su[SMELL/target]
  rain-GEN   wet   scent-POSS[3SG]
  "the wet smell of rain"
  (Kumcu 2021), see p. 247)

In the field of linguistics, Williams (1976) proposed a hierarchical model for linguistic synesthesia in English, as shown in Figure 1. However, Zhao *et al.* (2019) found three types of directionalities of linguistic synesthesia in Mandarin Chinese, including the unidirectional, bidirectional, and biased-directional transfers between the senses, as demonstrated in Figure 2. With respect to the neuro-cognitive characteristics of linguistic synesthesia, most linguistic studies on linguistic synesthesia considered it a type of metaphor (Shen 1997; Shen and Cohen 1998; Yu 2003; Popova 2005; Shen and Gil 2008; Strik Lievers 2017). However, Cacciari (2008) and Ramachandran and Hubbard (2001) highlighted the neuro-biological nature of linguistic synesthesia, where linguistic synesthesia was argued to pattern with neurological synesthesia constrained by "strong anatomical constraints" (Ramachandran and Hubbard 2001, see p. 18).[a] More recently, Zhao *et al.* (2022) have clarified linguistic synesthesia as a type of conceptual metaphor, where lexicalized concepts of sensory properties are involved, rather than the real-time sensory input that is processed as in neurological synesthesia. Thus, similar to metaphor, linguistic synesthesia is also one of the important vehicles, through which we can further our understanding of semantics and cognition.

However, compared to the studies on metaphor detection which have received notable results (Turney *et al.* 2011; Bulat, Clark, and Shutova 2017; Su, Wu, and Chen 2021), very limited research has yet been devoted to automatic linguistic synesthesia detection. Although linguistic synesthesia is similar to metaphor with conceptual mappings, linguistic synesthesia detection cannot directly apply the metaphor detection methods without significant modifications. That is, linguistic synesthesia involves conceptual mappings from one concrete sensory domain to another concrete sensory domain, while metaphor generally exhibits conceptual mappings from concrete domains to abstract domains (Zhao *et al.* 2022). Thus, there is a research gap in the computational

---

[a]Neurological synesthesia involves the association of perceptions in perceptual experiences whereby sensations in one sensory modality can be perceived when a different modality is stimulated (e.g., tasting shapes), or perception in one sub-modality can be obtained when another sub-modality is stimulated (e.g., perceiving colors from black-printed graphemes) (Cytowic 2002; Simner and Hubbard 2013)

analysis of linguistic synesthesia, where the task of automatic synesthesia detection has been given little attention in NLP. Jiang *et al.* (2022)'s study is the only exception, which aimed to detect linguistic synesthesia automatically in Mandarin Chinese through a radical-based neural model. However, the study is only a pilot study, in which only radical information was incorporated. In other words, the model proposed by Jiang *et al.* (2022) is language and writing system-dependent, which cannot be easily applied to other languages. Thus, this study proposes a neural network that leverages culturally enriched linguistic information including word segmentation and part-of-speech (POS) features, which can be generalized to other languages. Specifically, two kinds of linguistic features are utilized: the sub-lexical-level features including characters in the original text and the main semantic symbols of the characters (i.e., radicals), and the word-level features including segmented word sequences and their corresponding POS tags. Based on the extensive linguistic studies on linguistic synesthesia (Strik Lievers 2015; Winter 2019a; Zhao 2020), content words such as adjectives, verbs, nouns, and adverbs are frequently involved in linguistic synesthetic usages. In Chinese synesthesia particularly, the radical information in Chinese characters could provide important clues for determining the sensory modalities of words (Zhao, Huang, and Long 2018 Zhao, Huang, and Ahrens 2019; Zhao, Ahrens, and Huang 2022). Thus, this study presumes that both the character information and the word information would contribute to the neural model for the detection of linguistic synesthesia.

The task of linguistic synesthesia detection conducted by this study would contribute to both computational analyses and linguistic studies of the phenomenon in the following respects. Firstly, linguistic synesthesia involves sensory words and hence crucially reports the physical world as perceived by the speaker, which thus facilitates contextualizing NLP representations in the real world. More specifically, sensory information showing great usefulness in the task of sentiment analysis, has been illustrated in Picard (2000), Xiang *et al.* (2021), and Zolyomi and Snyder (2021). Thus, linguistic synesthesia encoding sensory information and showing regular patterns of sensory inputs would also show usefulness in the task of sentiment analysis. For example, Zhong *et al.* (2022) illustrated that the gustatory perceptions of 辣味 *la4wei4* "spicy taste" and 麻 *ma2* "numbing" were described most frequently in terms of linguistic synesthesia using words related to hurt and irritation, which were generally unpleasant (e.g., 燒嘴 *shao1zui3* "burning the mouth (unpleasantly spicy)"). Secondly, one promising application of linguistic synesthesia detection is concerned with the clinical pre-diagnosis for neurological synesthesia. That is, studies by Rizzo (1989), Cytowic (2002), and Turner and Littlemore (2023) showed that people who could experience neurological synesthesia generally employed peculiar linguistic synesthetic descriptions (e.g., "tasting the shape"). Thirdly, most of the existing studies on linguistic synesthesia rely on the extraction of synesthetic data manually or semi-automatically, which are time-consuming (Strik Lievers 2015; Zhao, Huang, and Ahrens 2019; Kumcu 2021). The automatic methods to detect linguistic synesthesia in natural language would improve the efficiency of data collection. Last but not least, the computational models for linguistic synesthesia leveraging linguistic features could test the correlations between linguistic features and the patterns of linguistic synesthesia attested by linguistic studies, based on the extent to which a specific linguistic feature can improve the performance of the models.

To summarize, this study aims to fill in the gap in automatic linguistic synesthesia detection. The main contributions of our work can be summarized as follows:

- This study proposes a neural network model that leverages culturally enriched linguistic information for linguistic synesthesia detection. As the word-level linguistic features employed are not language-specific, our model could be generalized for the detection of linguistic synesthesia in other languages.
- We construct a Chinese synesthesia dataset with rigorous annotations.

- Comprehensive experiments show that our model outperforms the state-of-the-art baseline models and achieves the best performance on the Chinese synesthesia dataset for the task of Chinese synesthesia detection.
- In addition to facilitating data collection, our model shows various potential applications, such as in the sentiment analysis, the clinical pre-diagnosis of neurological synesthesia, and linguistic theories about figurative languages.

In what follows, Section 2 reviews the related work on the detection of metaphor and linguistic synesthesia. Following that, a detailed description of the dataset and linguistic features is presented in Section 3. Section 4 and Section 5 focus on the proposed methods and the experiments respectively. Section 6 summarizes the results of this study. After that, the last section presents the limitations of this study and suggests future work.

2. Related work

2.1 Metaphor detection

Studies on the processing of metaphors have developed various models to automatically detect metaphorical expressions in a sentence. These studies can be divided into three categories based on the computational methods utilized: the feature-based approach, the shallow network-based approach, and the contextualized approach.

Regarding the feature-based approach, various linguistic features related to metaphorical expressions have been proposed and incorporated into (mostly) linear classifiers. The features (mainly in English) include word abstractness and concreteness (Turney *et al.* 2011), word imageability (Broadwell *et al.* 2013), semantic supersenses (Tsvetkov *et al.* 2014), and property norms (Bulat, Clark, and Shutova 2017). In Mandarin Chinese, radical information and sensory information were also employed (Chen *et al.* 2017, Wan *et al.* 2020). However, designing features based on human knowledge is expensive, and low-frequency metaphorical features are often neglected.

With the development of neural networks, several studies proposed neural metaphor detection models using recurrent neural networks (RNNs) or convolutional neural networks (CNNs). For instance, Wu *et al.* (2018) combined CNN and LSTM layers to utilize local and long-range contextual information to identify metaphorical details. In addition to the POS and word clustering information, Wu *et al.* (2018) also employed text information as linguistic features. Gao *et al.* (2018) showed that relatively standard BiLSTM models that operated on complete sentences worked well in the task of metaphor detection by formulating the task as sequence labeling or classification. These models outperform linear classification models by a significant margin and also avoid most of the feature annotation processes.

With respect to the contextualized approach, the contextualized language modeling coupled with a transformer network can encode semantic and contextual information. It can thus detect metaphors with fine-tuning training like other tasks. For instance, Su *et al.* (2021) introduced a variety of linguistic features (i.e., global/local text context and POS features) into the field of computational metaphor detection by leveraging powerful pre-training language models (i.e., RoBERTa). Gong *et al.* (2020) applied linguistic information from external resources such as WordNet with a similar RoBERTa network. Choi *et al.* (2021) proposed a metaphor-aware late interaction over the BERT (MelBERT) model, which leveraged the contextualized word representation and relevant linguistic metaphor identification theories to detect whether the target word is metaphorical.

To summarize, the different approaches for metaphor detection vary in their computational models. However, linguistic features are generally incorporated into the models, which show great contributions to the improvements of the performances of the models on the detection tasks.

2.2 Linguistic synesthesia detection

Different from extensive work on metaphor detection, there have been only several studies reported to focus on the detection of linguistic synesthesia in natural language. These studies can be classified into two categories: one is to employ semi-automatic methods, and the other is to utilize automatic methods based on neural models. Strik Lievers *et al.* (2013) and Strik Lievers and Huang (2016) proposed a semi-automatic approach to extract synesthetic expressions in English and Italian. The approach needed a lot of manual strategies, such as compiling a list of perception-related lexical items and manually selecting sentences that contained linguistic synesthesia. Following a similar method to Strik Lievers *et al.* (2013), Liu *et al.* (2015) extracted linguistic synesthetic sentences for Mandarin Chinese. These semi-automatic approaches are expensive and time-consuming.

With respect to detecting linguistic synesthesia via neural networks, a recent work by Jiang *et al.* (2022) is the first to propose the task of Chinese synesthesia detection. The study provided a family of baseline models for linguistic synesthesia detection. In addition, a radical-based neural model was proposed for linguistic synesthesia detection. However, there have been notable limitations of the work by Jiang *et al.* (2022) in the linguistic feature selection, the data annotation, and the experiment design. From the feature engineering perspective, the study only incorporated the radical information of Chinese characters as the linguistic feature into the model. Thus, Jiang *et al.* (2022)'s model is language and writing system-dependent and cannot be easily generalized to other languages. In addition, the orthographic information of the radical components in Chinese orthography was not utilized appropriately by Jiang *et al.* (2022). That is, Jiang *et al.* (2022) relied on the Xinhua dictionary which was designed for simplified Chinese characters to detect radical information, while the dataset utilized by the study contained linguistic expressions in traditional Chinese characters. In terms of the annotation process of linguistic synesthetic data by Jiang *et al.* (2022), the annotators were not given rigorous training on how to decide linguistic synesthetic usages before the annotation, except being provided with written instructions. With respect to the experiment process, Jiang *et al.* (2022) used the golden label of sensory word extraction as the input of sensory modality detection. However, a boundary detection model is generally used first to detect the sensory word boundary.

This study leverages culturally enriched linguistic features for the automatic detection of linguistic synesthesia. Specifically, apart from the radical information, the word segmentation and POS features are incorporated, to ensure that the proposed model could be applied to other languages. In addition, a more compatible and conventionalized conceptual orthographic system for Chinese traditional characters (i.e., Hantology[b]) is utilized to detect radicals.[c] Furthermore, a linguist is invited to give a detailed introduction to linguistic synesthesia, to ensure that annotators have sufficient knowledge about the phenomenon before annotation. Last but not least, this study refines the experiment setting by adopting a boundary detection model for word identification (Huang *et al.* 2007) rather than employing the golden labels (i.e., the sensory words annotated in the dataset) for linguistic synesthesia detection.

---

[b]Hantology (Hanzi Ontology) is a language resource designed to contain three-level information for Chinese characters: meanings of the characters, radicals of the characters, and meaning mappings to SUMO (The Suggested Upper Merged Ontology) (Chou and Huang 2006, 2010), which can be accessed at: https://hantology.ling.sinica.edu.tw. For the application of Hantology in the NLP analysis of Chinese, please refer to Chen *et al.* (2017, 2019) on metaphor detection and Chen *et al.* (2021) on emotion classification.

[c]As stated in Section 1, based on the previous studies on linguistic synesthesia, both the character information and the word information would contribute to the neural model for the detection of linguistic synesthesia. The details of the relevance of each feature for the detection task will be given in Section 3.2.
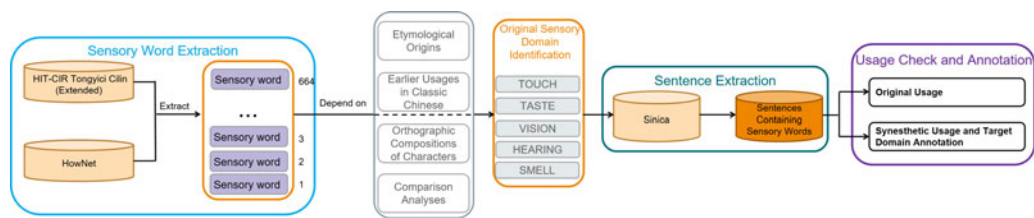
Figure 3. The procedure for dataset acquisition.

3. Dataset and linguistic features

3.1 Dataset

This study followed a linguistic synesthesia identification procedure proposed by Zhao (2020) to construct the dataset, which is adapted as Figure 3. Although no consensus has been reached regarding the classification of human senses (Miller and Johnson-Laird 1976; Purves *et al.* 2001), the Aristotelian five senses (i.e., touch, taste, vision, hearing, and smell) have generally been utilized to analyze linguistic synesthesia (Strik Lievers 2015; Zhao, Huang, and Long 2018; Winter 2019a; Winter 2019b; Kumcu 2021). Based on the five sensory modalities, 664 sensory words were extracted automatically from two Chinese lexical thesauri, including HIT-CIR Tongyici Cilin (Extended) (Che, Li, and Liu 2010) and HowNet (Dong and Dong 2003). In order to identify the original sensory domain of each of the 664 sensory words, the etymological origins, the earlier usages in Classic Chinese, the orthographic compositions of characters, and the comparison analyses were employed collectively. The step was conducted by a Chinese linguist. For example, the sensory adjective 濃 *nong2* has two different orthographic writings in Classic Chinese: one is 醲 with the radical denoting wine, which was used to describe the strong taste of wine; and the other is 濃 with the radical denoting water, which was used to describe the visual sensation of dense dew (Xu 156; Duan 1815). Thus, it is not easy to decide which sensory modality (i.e., taste or vision) is the original domain for the adjective 濃 *nong2*. However, the adjective 濃 *nong2* was used most frequently to show a comparison with the adjective 淡 *dan4* in Classic Chinese, whose original meaning was paraphrased as "mild taste" in Chinese dictionaries (Xu 156; Duan 1815). Thus, the comparison analysis demonstrated that taste was the most likely to be the original sensory domain of the adjective 濃 *nong2* as well.

After determining the original sensory domain of the sensory words, the sentences containing the words were extracted from the Sinica corpus (Chen *et al.* 1996).[d] Three undergraduate students were trained to decide whether the usages of the sensory words were synesthetic before the annotation. Then, we asked the three annotators to manually check whether the usages of the sensory words belonged to the original sensory domains of the words: if yes, the usages were marked as original usages; if not and the usages still described sensory perceptions, the usages were marked as synesthetic usages. For the synesthetic usages, the target domains of the sensory words were also annotated. Figure 4 shows an example of annotation for linguistic synesthesia in Mandarin Chinese. Specifically, the sensory adjective 冰冷 *bing1leng3* "cold" has its original sensory modality as touch. However, the adjective was used to describe hearing in the expression 一個冰冷憤怒的聲音 *yi1ge4 bing1leng3 fen4nu4 de sheng1yin1* "a cold and angry voice". Thus, the sensory word 冰冷 *bing1leng3* "cold" was marked with the linguistic synesthetic usage, and its original and target modalities were annotated as touch and hearing respectively.

Through the annotation process, 187 sensory adjectives with both original and synesthetic usages were identified, where 7,217 synesthetic sentences were annotated. Table 1 demonstrates

---

[d]The Sinica corpus (The Academia Sinica Balanced Corpus) can be accessed at: http://lingcorpus.iis.sinica.edu.tw/modern/.

Table 1. Inter-annotator agreements for annotation of linguistic synesthesia

	Annotation	Kappa score
Synesthesia Annotation	Synesthetic usage	0.835
Domain Annotation	TOUCH as the target	0.779
	TASTE as the target	0.854
	VISION as the target	0.862
	HEARING as the target	0.884
	SMELL as the target	0.832

Table 2. Data distribution of the five sensory modalities in synesthetic and original sub-datasets

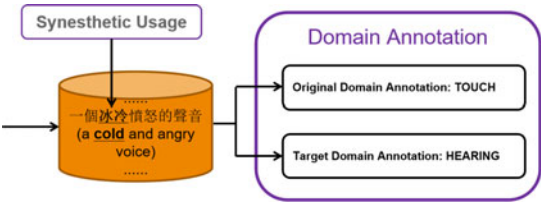|	Synesthetic sub-dataset		Original sub-dataset	
Sensory modality	Type	Token	Type	Token
TOUCH	69	2,361	69	2,361
TASTE	20	2,097	20	2,097
VISION	92	2,697	92	2,697
HEARING	4	33	4	33
SMELL	2	29	2	29
Total	187	7,217	187	7,217

Figure 4. An example of annotation of linguistic synesthesia in Chinese.

the inter-annotator agreements measured by the kappa scores (Fleiss 1971). Our annotation has reached kappa scores of 0.779 to 0.884, which are all higher than that by Jiang *et al*. (2022) (i.e., 0.757).

As original usages are generally more frequent than synesthetic usages for Mandarin sensory adjectives (Zhao, Ahrens, and Huang 2022), 7,217 original usages were randomly extracted for the 187 sensory adjectives, in accordance with the distribution of the five sensory modalities in the collected linguistic synesthetic data. Thus, both synesthetic sub-dataset and original sub-dataset were constructed, with the data distribution demonstrated in Table 2.[e]

---

[e]The whole dataset can be accessed at: https://osf.io/73tyc/?view_only=b5a503ce329948f2a3d73fa67ffb26a8.

3.2 Linguistic features

Linguistic features show significant usefulness in improving the performance of computational models for automatic metaphor detection, as reviewed above. This current study incorporates linguistic features including character information, word segmentation, and POS features into the neural network model to detect linguistic synesthesia automatically in Mandarin Chinese.

Character features. Although the character is generally regarded as an orthographic unit, it can also act as an important syntactic and semantic unit in Chinese (Xu 2005; Ye 2015). With respect to the NLP tasks, Chen *et al.* (2017, 2019), Hou *et al.* (2019), and Chen *et al.* (2021) improved the performances of the computational models by introducing the Chinese character as an independent linguistic feature on metaphor detection, register classification, and emotion classification respectively. The character is also an important linguistic feature for linguistic synesthesia detection, whose radical component provides a conventionalized clue for detecting the original sensory domain of the lexical item represented by the character. Woon and Yun (1987) found that over 80 percent of Chinese characters were phono-semantic compounds, where a semantic component (mostly a radical) indicated a broad category of the meaning of the character. For instance, the radical of 冷 *leng3* "cold" means ice, which indicates touch as the original domain for the adjective. Similarly, the radical of 甜 *tian2* "sweet" is 舌 denoting the tongue, through which humans experience gustatory perceptions. Thus, the original sensory domain of 甜 *tian2* "sweet" is taste. In addition, there are abundant sensory words with synesthetic usages in Mandarin, which only contain one single character. Examples are such as 冷 *leng3* "cold" in 冷香 *leng3 xiang1* "cold fragrance" and 甜 *tian2* "sweet" in 甜白 *tian2 bai2* "sweet white".

Word segmentation features. Compounding is a productive morphological device for word formation in Mandarin Chinese (Chao 1968; Huang and Shi 2016). Mandarin compound adjectives can also be involved in linguistic synesthesia (Zhao 2020; Zhao, Ahrens, and Huang 2022). For instance, the monosyllabic visual word 大 *da4* "big" can be duplicated as one single word 大大 *da4da4* "big" to describe hearing as in 大大的聲音 *da4da4 de sheng1yin1* "a big sound". Besides, two different monosyllabic words can be combined as a compound word used for linguistic synesthesia. For instance, the monosyllabic word 甜 *tian2* "sweet" and the monosyllabic word 美 *mei3* "tasty" can be compounded as a single word 甜美 *tian2mei3* "tasty". The compound word 甜美 *tian2mei3* "tasty" with taste as its original domain, can be used in linguistic synesthesia for vision, as in 甜美 的長相 *tian2mei3 de zhang3xiang4* "a sweet appearance". As one sub-task of this study is to detect sensory words with linguistic synesthetic usages, the word segmentation information would be of great usefulness to detect the boundary of the sensory words.

POS features. Linguistic synesthesia was found to show certain patterns on syntactic structures (Strik Lievers 2015; Zhao, Ahrens, and Huang 2022). For instance, typical synesthetic expressions in English and Italian are composed of a sensory adjective acting as the source and a noun as the target (Strik Lievers 2015). Linguistic synesthetic expressions in Mandarin frequently exhibit the syntactic combinations of "adjective + noun" (e.g., 喧鬧的色彩 *xuan1nao4 de se4cai3* "a loud color"), "adverb + verb" (e.g., 重重地說 *zhong4zhong4 de shuo1* "saying in a heavy voice"), and "verb + noun" (e.g., 聞到花香 *wen2 dao4 hua1xiang1* "to smell the fragrance of flowers") (Zhao 2020). On the contrary, function words (e.g., pronoun, preposition, conjunction, interjection, etc.) have not yet been reported to show linguistic synesthetic usages. Thus, the POS information would contribute to the computational models for linguistic synesthesia detection.

4. Proposed methods

There are mainly three challenges in the automatic detection of linguistic synesthesia. Firstly, the target modality of one sensory word may vary in its different contexts. For instance, the tactile adjective 尖銳 *jian1rui4* "sharp" has its target modality as vision when used for 地形 *di4xing2*
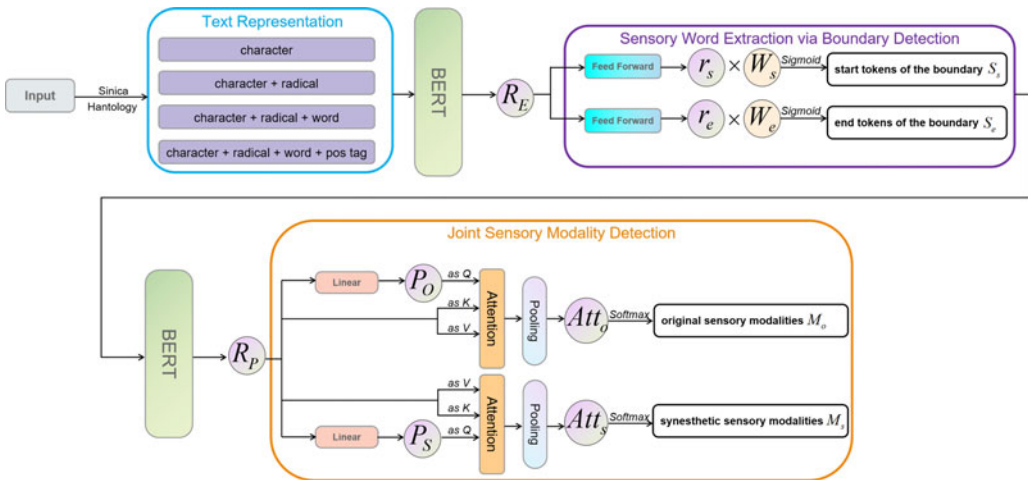
Figure 5. The architecture of our proposed methods.

"terrain" while as hearing in the context of the description of 聲音 *sheng1yin1* "sound". Hence, it is necessary to capture both the sensory expressions of the sensory word and its contexts. Secondly, one sensory word may not contain a single character, as Mandarin compounds can also be used for linguistic synesthesia (see Section 3.2). Thus, it is necessary to detect the boundary of the sensory word. Thirdly, there is an association between original and synesthetic sensory modalities. For instance, taste is significantly correlated with smell, and vision is significantly associated with hearing in Mandarin synesthesia (Zhao 2020). It therefore makes modeling the interaction between sensory modalities necessary.

This study proposes a multi-linguistic feature-based end-to-end neural model to address the three challenges, with the overall architecture of the methods shown in Figure 5. Specifically, the linguistic features include both sub-lexical and word-level features. The sub-lexical-level feature includes characters in the original text and the main semantic symbols of the characters (i.e., radicals) obtained from an external knowledge base (i.e., Hantology). The word-level information includes segmented word sequences and their corresponding POS tags, which can be obtained directly from the Sinica corpus or by using the existing Chinese word segmentation/POS tagging system like Jieba tokenizer[f] or Stanford CoreNLP.[g] For modeling the sensory word extraction and linguistic synesthesia detection simultaneously, our model includes the following three steps:

- **Text representation:** building multi-linguistic features based on the text representation and using different features to capture the relationship between sensory words and their contexts.
- **Sensory word extraction via boundary detection:** extracting sensory words based on a span-based boundary detection model.
- **Joint sensory modality detection:** predicting the original sensory modality of the sensory words and classifying the actual sensory modality (i.e., the synesthetic sensory modality) in the text collectively, based on the sensory words extracted in the previous steps and their contexts.

---

[f]https://github.com/fxsjy/jieba
[g]https://github.com/elisa-aleman/StanfordCoreNLP_Chinese

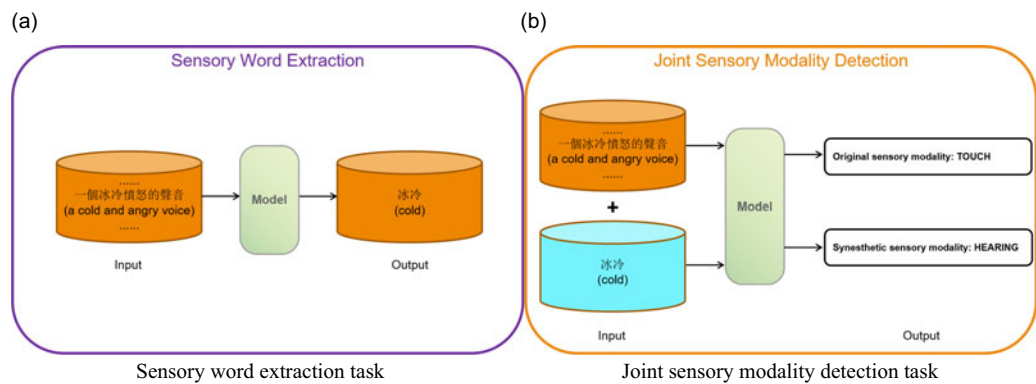Sensory word extraction task   Joint sensory modality detection task

Figure 6. An example for two sub-tasks: sensory word extraction and joint sensory modality detection.

4.1 Task definition

Given a raw text of Chinese characters $C = \{c_1, c_2, \ldots, c_n\}$, the desired output contains the sensory word and its related sensory modalities from both the original and target domains. Thus, the linguistic synesthesia detection task is divided into a two-part model with two sub-tasks: **sensory word extraction** and **joint sensory modality detection** (see Figure 5)

- **Sensory word extraction:** its goal is to extract the sensory word in the given raw text, modeled as the sequence labeling task. As shown in the annotated example in Figure 6(a), the input of this sub-task is the raw Chinese sequence C, and the output of this sub-task is the sensory word 冰冷 *bing1leng3* "cold," since the word is used in linguistic synesthesia for hearing rather than touch in the given text.

- **Joint sensory modality detection:** its goal is to predict the original and synesthetic sensory modalities of the previously extracted sensory word, modeled as the text classification task. As shown in the annotated example in Figure 6(b), the input consists of the raw Chinese sequence C and the extracted sensory word from the previous sub-task, and the output is the original sensory modality as touch and the synesthetic sensory modality as hearing for the word 冰冷 *bing1leng3* "cold."

4.2 Text representation

In the process of building the text representation, the model uses four different text representation methods for the original text, namely "character", "character + radical", "character + radical + word", and "character + radical + word + pos tag". Among them, the radical is the part of the Chinese character that specifies the meaning category. For example, the main radical of the Chinese character 吃 *chi1* "eat" is 口 "mouth". Therefore, we integrate radicals into the text representation. In addition, the word segmentation and the POS information used in this model are based on the original annotations in the Sinica corpus. Formally, given a Chinese raw text C, it contains n characters, i.e., $C = \{c_1, c_2, \ldots, c_n\}$, where each character c_i is an independent item. Then, the characters are mapped into the radicals respectively by looking up Hantology, i.e., $H = \{h_1, h_2, \ldots, h_m\}$. As to the word segmentation information in the Sinica corpus, we convert the original text C into a word sequence of m length as $W = \{w_1, w_2, .., w_m\}$, and the corresponding POS sequence is $P = \{p_1, p_2, \ldots, p_m\}$.

Our model includes two parts, i.e., the sensory word extraction and the sensory modality detection. However, the explicit information in the sensory modality detection task contains the sensory word e obtained from the sensory word extraction task, so the textual representations of the two

parts are not the same. We thus utilize BERT (Devlin *et al.* 2019) to learn the representation R_E for the sensory word extraction and R_P for the sensory modality detection, with the details as:

- **Character** is the basic token-level information of the raw input. We use "[SEP]" token to separate the characters C and the extracted sensory word e for R_P to notify BERT that the sensory word has distinct significance when compared to other characters in the C. The text representations under the "character" feature are formulated as follows:
  $R_E = \text{BERT}([CLS]C[SEP]), R_P = \text{BERT}([CLS]C[SEP]e)$

- **Character + Radical** consists of token-level characters C and radical information H in the input. We use "[SEP]" to separate the characters, radical information, and the extracted sensory word. This approach enables the integration of radical information within the encoded representation, as each character is mapped to its respective radical in a one-to-one manner. The text representations under the "character + radical" feature are formulated as follows:
  $R_E = \text{BERT}([CLS]C[SEP]H), R_P = \text{BERT}([CLS]C[SEP]H[SEP]e)$

- **Character + Radical + Word** also incorporates word-level segmentation due to the necessity of tokenization in the Chinese language, such as word ambiguity. We also use "[SEP]" to divide different parts of the input. The text representations under the "character + radical + word" feature are formulated as follows:
  $R_E = \text{BERT}([CLS]C[SEP]H[SEP]W), R_P = \text{BERT}([CLS]C[SEP]H[SEP]W[SEP]e)$

- **Character + Radical + Word + POS tag** also focuses on word-level features, because sensory words are generally used as adjectives or adverbs in linguistic synesthetic usages. Similarly, we use "[SEP]" token to concatenate the different kinds of features in the encoder input. The text representations under the "character + radical + word + pos tag" feature are formulated as follows:
  $R_E = \text{BERT}([CLS]C[SEP]H[SEP]W[SEP]P), R_P = \text{BERT}([CLS]C[SEP]H[SEP]W[SEP]P[SEP]e)$

Note that "[CLS]" and "[SEP]" are more than *ad hoc* feature-marking tokens initiated in the pre-training procedure of BERT. The token "[CLS]" (classification) is the classification result of the entire sentence, and its hidden vector is influenced by all other words in the sentence. On the other hand, the token "[SEP]" (separation) instantiates the boundary between lexical units in a sentence. Huang *et al.* (2007) proposed the boundary detection model for word segmentation, and Li *et al.* (2012) showed that boundary detection was much more efficient and required less training data than word identification. Modeling these two concepts explicitly allows us to fully leverage the contextual information in BERT. By leveraging BERT's multi-head attention mechanism and pre-trained knowledge, each attention head learns unique patterns and relationships from various parts of the given input. This helps us to create linguistically enriched text representations that capture the deep connections between tokens and linguistic features.

4.3 Sensory word extraction via boundary detection

We then propose a boundary detection model to detect the boundary of sensory words. Therefore, the sensory word extraction is reformulated as the task of identifying the start and end indices of the sensory word (Hu *et al.* 2019; Wang *et al.* 2019). Given a sequence R_E from the text representation, we apply two separate feed-forward neural networks to create different representations (r_s/r_e) for the start/end of the spans. A sigmoid is introduced to produce the probability of each token being selected as the start/end of the scope:

$$r_s = \text{FFNN}(W_s \cdot R_E + B_s)$$

$$S_s = \text{Sigmoid}(r_s) \tag{1}$$

$$r_e = \text{FFNN}(W_e \cdot R_E + B_e)$$

$$S_e = \text{Sigmoid}(r_e) \tag{2}$$

where W_s, W_e and B_s, B_e are weights and biases in the model parameters, and S_s and S_e are the outputs of the sensory word extraction model, which are used to predict the start and end tokens of the boundary of the extracted sensory word.

4.4 Joint sensory modality detection

Given the text representation R_P and the sensory word extraction learned from the previous subsection, we propose a joint model to detect the sensory word's original and synesthetic sensory modalities simultaneously. Specifically, we first feed R_P into two separate feed-forward neural networks and obtain two representations, i.e., (r_o/r_a) for the original modality and the actual synesthetic modality, respectively:

$$r_o = \text{FFNN}\,(W_o \cdot R_P + B_o)$$

$$r_a = \text{FFNN}\,(W_a \cdot R_P + B_a) \tag{3}$$

where W_o, W_a and B_o, B_a are weights and biases in the model parameters. We then employ distinct attention layers to capture the relationship between different sensory modalities and the original texts and leverage attention to enhance the performance of the model by emphasizing key input elements, thereby improving accuracy in our task.

$$att_o = \text{Softmax}\left(\frac{r_o \cdot R_P{}^T}{\sqrt{d_{R_P}}}\right) R_P$$

$$att_a = \text{Softmax}\left(\frac{r_a \cdot R_P{}^T}{\sqrt{d_{R_P}}}\right) R_P \tag{4}$$

where d_{R_P} is the dimension of the representation R_P. After obtaining the hidden representation (att_o/att_a), we use two softmax layers to predict the original and synesthetic sensory modalities as follows:

$$M_o = \text{Softmax}([r_o; att_o]) \tag{5}$$

$$M_a = \text{Softmax}([r_s; att_a]) \tag{6}$$

where r_o is concatenated with att_o, and r_s is concatenated with att_a. M_o and M_a are the outputs of the sub-task, corresponding to the predicted original modality and the predicted actual synesthetic modality, respectively.

An example is shown in Table 3, to illustrate how linguistic features are represented and contribute to the detection of linguistic synesthesia. Specifically, the radical 耳 "ear" of the character 聲 *sheng1* "sound" is closely related to hearing, while the radical 金 "metal generally used for making weapons in ancient China" suggests touch as the most relevant sensory modality. The inconsistency of the sensory modalities (hearing vs. touch) within a sentence indicates a high likelihood of linguistic synesthetic usages. In addition, the word segmentation helps our model to identify the two characters 鈍鈍 *dun4dun4* "blunt" as a word, rather than the single character 鈍 *dun4* "blunt". Besides, the POS information is employed to detect 鈍鈍 *dun4dun4* "blunt" as the synesthetic word, as the sensory adjective involves linguistic synesthesia most frequently. Based on the comprehensive linguistic information of the character, radical, word segmentation, and POS features, the sensory word 鈍鈍 *dun4dun4* "blunt" with linguistic synesthetic usages is identified, whose original domain as touch and target domain as hearing are detected jointly, as shown in Figure 7.

Table 3. An example for representation of linguistic features[h]

Input	Plain text	聲音也還是鈍鈍的。 "The sound is still blunt."
Linguistic Features	Key characters and radicals	聲: 耳 [ear: HEARING]; 鈍: 金 [metal generally used for making weapons: TOUCH]
	Word segmentation + pos tag	聲音/Na 也/D 還/D 是/SHI 鈍鈍/VH 的/T。
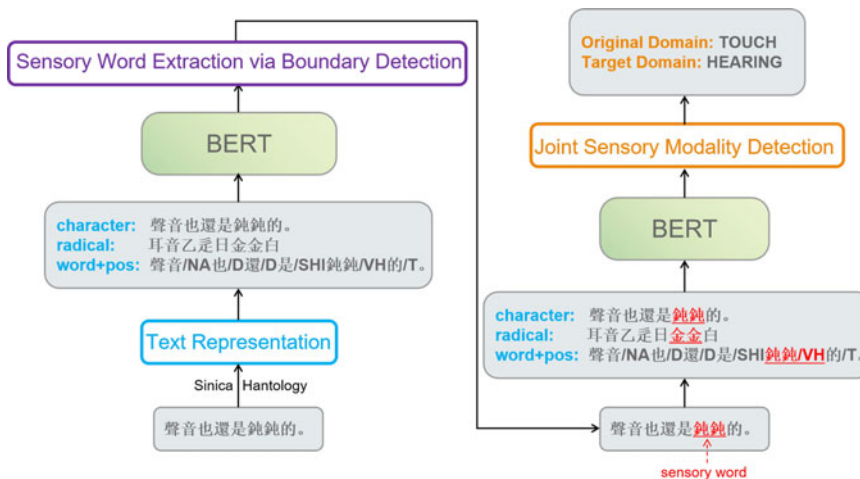
Figure 7. An example of linguistic synesthesia detection.

4.5 Training

We train the sensory modality detection with the sensory word extraction in a unified architecture.

Loss of the sensory word extraction. We minimize the negative log-likelihood loss to train the sensory word extraction model, and parameters are updated during the training process. In particular, the loss is the sum of two parts: the start token loss and the end token loss,

$$\mathcal{L}_S = -\sum y_s \log(S_s) - \sum y_e \log(S_e) \tag{7}$$

where y_s and y_e are the ground truth start and end positions for the sensory word extraction model.

Loss of the sensory modality detection. Our training objective of the sensory modality detection is to minimize the cross-entropy loss with a *l*2-regularization term,

$$\mathcal{L}_P = -\sum y_o \log M_o - \sum y_a \log M_a + \frac{\lambda}{2}||\theta_y||^2 \tag{8}$$

where y_o and y_a are the pre-defined labels for the original and actual sensory modalities, respectively. And λ is a parameter for *l*2-regularization.

Therefore, the final loss is demonstrated as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_S + \lambda_2 \mathcal{L}_P \tag{9}$$

where λ_1 and λ_2 are the trainable weight parameters, and $\lambda_1 + \lambda_2 = 1$.

---

[h]/Na, /D, /SHI, /VH, and/T are the POS tags based on the Sinica corpus, referring to the common noun, the adverbial, Chinese special verb 是 "BE", the stative intransitive verb, and the particle respectively.

5. Experiments

5.1 Experiment settings

One challenge in linguistic synesthesia detection is to find a large-scale dataset that includes rich linguistic synesthetic usages. Based on our collected Mandarin synesthesia dataset, the extraction and detection tasks both have 11,484 sentences in the training set, 1,460 sentences in the testing set, and 1,456 sentences in the development set (roughly following the 8:1:1 ratio). We have double-checked that there is no overlap between the testing dataset used in the sensory word extraction task and the testing dataset in the sensory modality detection task. In addition, the sensory words in the training set did not appear in the testing set.

This study used the BERT-base-Chinese as our pre-trained model. The optimizer chosen is Adam (Kingma and Ba 2015), and the parameters of BERT and other models are optimized separately. Besides, this study utilized a lower learning rate of 1e-5 with a training batch size of 16. For LSTM-based baselines, we used the 50-dimensional character embeddings, which were pre-trained on Chinese Giga-Word using Skip-gram word2vec (Mikolov *et al.* 2013) and fine-tuned during the model training. All experiments were conducted on a single NVIDIA GeForce RTX 1080 Ti (11 GB memory). It is important to note that different from Jiang *et al.* (2022) using the golden data of the sensory word for linguistic synesthesia detection, this study considered the pipeline setting and used the prediction results of sensory word extraction for linguistic synesthesia detection. The selected evaluation metrics (i.e., Precision, Recall, F1 score) were calculated via the scikit-learn and SeqEval[i] packages.

5.2 Baseline selection

The task of sensory word extraction aims to extract the perception-related word from a sentence. Generally speaking, it can be considered a sequence labeling task. On the other hand, linguistic synesthesia detection aims to detect the original and synesthetic sensory modalities of the given sensory word. Therefore, this task can be separated into two sub-tasks: original sensory modality detection and synesthetic sensory modality detection. These two sub-tasks can be considered two text classification tasks.

Firstly, our work chooses the following baselines for both the sensory word extraction task and the sensory modality detection task:

- E2EB-iLSTM: a relatively standard end-to-end BiLSTM model for detecting the metaphorical use of words in context proposed by Gao *et al.* (2018).
- MelBERT: originally developed for the metaphor detection task, namely the metaphor-aware late interaction over BERT (i.e., MelBERT) (Choi *et al.* 2021). The model leveraged the contextualized word representation and linguistic features to detect whether the target word is metaphorical.

Then, the following models are developed by this study as baselines only for the task of sensory word extraction:

- BiLSTM + CRF: as BiLSTM + CRF (Lample *et al.* 2016) is widely used in many sequence labeling tasks, we adopt it as an essential baseline for sensory word extraction. In particular, we apply a bidirectional LSTM (Schuster and Paliwal 1997) as the textual encoder and the conditional random fields (CRF) (Lafferty, McCallum, and Pereira 2001) as the decoder.

---

[i]https://pypi.org/project/seqeval/

- BERT + CRF: instead of training a model from scratch, we also adopt the framework of fine-tuning a pre-trained language model on a downstream task (Radford *et al.* 2018). In this framework, we adopt BERT (Devlin *et al.* 2019) as the textual encoder and use CRF as the decoder.

- BERT + MRC: it is the same pre-training and fine-tuning model as BERT + CRF. Instead of CRF as the decoder, we formulate it as a machine reading comprehension (MRC) task (Chen and Wu 2020). Specifically, we first utilize the original raw text as the input passage to the BERT encoder. Then, we follow Li *et al.* (2020) to employ two separate feed-forward layers over the text representation to generate distinct representations for the start and end of the spans. This allows our model to effectively identify and locate the relevant spans of the sensory word.

In addition, for sensory modality detection, we select several state-of-the-art models in metaphor detection, aspect-based sentiment analysis, and other related text classification tasks:

- SR-BiLSTM: similar to the standard LSTM model struggling to detect the important part for metaphor detection, SR-BiLSTM (Sensory-Related BiLSTM) is implemented by our study based on a minor modification of the TD-LSTM model originally designed for aspect-based sentiment analysis (Tang *et al.* 2016). The baseline model uses an attention mechanism that can capture the critical part of a sentence in response to a sensory word (Wang *et al.* 2016) and a bidirectional LSTM (Schuster and Paliwal 1997) as the encoder of the sensory word and the content of the sentence. Then, SR-BiLSTM employs an attention mechanism to explore the connection between the sensory word and the content.

- PF-BERT: due to the importance of the context of the sensory word in linguistic synesthesia detection, we model the preceding and the following contexts surrounding the sensory word. Therefore, contexts in both directions can be used as feature representations for synesthesia detection. In particular, we build a baseline model called PF-BERT (Preceding and Following BERT), which uses two BERT neural networks (Devlin *et al.* 2019) to model the preceding and the following contexts respectively.

- MrBERT: the metaphor-relation BERT model (MrBERT) explicitly models the relationship between a verb and its grammatical, sentential, and semantic contexts (Song *et al.* 2021). The model is employed to frame sensory modality detection as a relation extraction task, which enables modeling the synesthetic relation between a sensory word and its context components and uses the relation representation for determining linguistic synesthesia of the word.

The baseline models mentioned above can be roughly divided into two groups based on the type of encoder they use: LSTM-based models and BERT-based models. The primary difference between LSTM-based and BERT-based models lies in the design and learning technique. LSTM is a traditional neural network for sequential data that focuses on short-term dependencies. BERT, on the other hand, is a trending and powerful large language model that has been pre-trained on a large amount of data and knowledge, allowing it to capture long-term dependencies and complex contextual information. Furthermore, the models only for extraction tasks can also be roughly categorized into the CRF-based model and the MRC-based model in terms of the type of decoder they utilize. After obtaining the text representation from the encoder, the CRF-based model will calculate the conditional probabilities of the output sequence, taking into account label dependencies. In contrast, the MRC-based model will focus on detecting the span boundary of the required output, aiming to identify the relevant answer within the given passage.

Table 4. The results of sensory word extraction

	Method	F1
LSTM-based	BiLSTM+CRF	71.5
	E2EB-iLSTM	74.6
BERT-based	BERT + CRF	77.9
	BERT + MRC	78.3
	MelBERT	80.2
Ours	character	78.4
	character+radical	**81.3**
	character+radical+word	81.0
	character+radical+word+pos tag	81.1

5.3 Results and discussion

This section presents the results of experiments for both the task of sensory word extraction and the task of joint sensory modality detection. After that, we show an analysis of factors that contribute to the performance of our model.[j]

5.3.1 Sensory word extraction

As shown in Table 4, the transformer-based models outperform the LSTM-based models for sensory word extraction. Specifically, our models and the three selected BERT-based baselines achieve more than 3 points higher than the LSTM-based models in the F1 score. These results show the effectiveness of BERT-based models for learning the sentence representation for sensory word extraction, regardless of whether the decoder is CRF or MRC. On the other hand, comparing the results of the BiLSTM+CRF model with the BERT + CRF model reveals that contextual information is quite useful in CRF-based extraction tasks. In addition, compared to the BERT + CRF model, the BERT + MRC model performs better, which shows that the boundary detection-based model is more effective than the traditional sequence labeling model. Thirdly, our proposed model achieves a state-of-the-art result (with the F1 score of 81.1) and outperforms other baseline models significantly ($p < 0.05$).

In terms of the four different linguistic features leveraged, the usage of Hantology contributes to the task of sensory word extraction significantly, resulting in a 2.9 percent improvement over the character-only model. However, adding the word segmentation and POS tagging features into the radical-incorporated model does not show an improvement in the performance of the model. These results prove that the task of sensory word extraction is sensitive to the sub-lexical-level knowledge that specifies the semantic and cognitive category of what is denoted by the character. These results also echo the work by Chen *et al.* (2017), which leveraged the radical information for the classification of ontological categories to improve the performance of the neural model for metaphor detection.

The results that character and radical features play a crucial role in the task of sensory word extraction have two important implications for the research on linguistic synesthesia. One is that linguistic synesthesia is the most likely to be involved in the monosyllable word composed of one character. In fact, Chen *et al.* (2019)'s experimental study and Zhao (2020)'s corpus-based

---

[j]All the experimental results in this section represent the average of ten independent runs. We conducted t-tests to assess the differences between our model and comparison models, consistently yielding a *p*-value smaller than 0.01 across all comparisons. This statistical significance underscores the robustness and superiority of our model's performance.

Table 5. The results of original modality detection, with F1 (weighted F1) calculated by taking the mean of all per-class F1 scores while considering the weight of each class

Method	Original					
	TOUCH	TASTE	VISION	HEARING	SMELL	F1
SR-BiLSTM	48.4	48.3	40.5	11.8	0.0	44.5
E2EB-iLSTM	45.4	49.3	57.2	0.0	0.0	50.2
PF-BERT	60.3	51.7	65.8	**53.3**	0.0	58.9
MelBERT	54.0	58.2	63.2	33.3	0.0	57.7
MrBERT	50.6	61.1	62.7	33.3	0.0	57.2
Ours(character)	60.1	58.5	60.2	27.3	0.0	58.6
Ours(character+radical)	**61.0**	**65.8**	63.6	33.3	0.0	62.3
Ours(character+radical+word)	55.6	63.8	64.0	46.2	5.1	60.2
Ours(character+radical+word+pos tag)	60.9	63.1	**66.1**	40.0	**6.7**	**62.5**

study also found a great numerical advantage of monosyllabic words (i.e., the word containing one character) with linguistic synesthetic usages over compounding words (i.e., the word containing more than one character). Thus, with respect to the sensory word extraction task for linguistic synesthesia, the character boundary and the word boundary overlap in most cases. That may be the reason why adding the word segmentation information does not improve the performance of the model which has already incorporated the character information for sensory word extraction. The other implication for studying linguistic synesthesia is that radical components of Chinese characters conceptualize comprehensive and systematic sensory information that may imply a culturally grounded conceptualization of semantics and cognition. For example, radicals denoting instruments (e.g., 火) are generally related to touch (e.g., 熱), radicals denoting the tongue (e.g., 舌) are generally related to taste (e.g., 甜), radicals denoting the nose (e.g., 自) are generally related to smell (e.g., 臭), radicals denoting the light (e.g., 日) are generally related to vision (e.g., 暗), and radicals denoting the mouth (e.g., 口) are generally related to hearing (e.g., 吵).[k]

5.3.2 Sensory modality detection

In terms of the task of sensory modality detection, our proposed model is compared with several classification baseline models in Tables 5 and 6, where SR-BiLSTM, E2EB-iLSTM, PF-BERT, MelBERT, and MrBERT are all state-of-the-art models for metaphor detection.

Based on the results in Tables 5 and 6, we find that:

- The performances of detection of the synesthetic modalities largely surpass those of detection of the original modalities in all the models.
- Our proposed model outperforms other baseline models significantly ($p < 0.05$) and reaches acceptable results in both the original modality detection and the synesthetic modality detection. The results indicate that leveraging linguistic features and joint learning is effective in linguistic synesthesia detection. With respect to the four features leveraged, the model containing all the features (i.e., "character + radical + word + pos

---

[k]Please note that although the mouth can also be used for tasting in perceptions, Chinese characters with the radical denoting the mouth are predominantly related to hearing (see the appendices of Zhao (2020)).

Table 6. The results of synesthetic modality detection, with F1 (weighted F1) calculated by taking the mean of all per-class F1 scores while considering the weight of each class

| Method	TOUCH	TASTE	VISION	HEARING	SMELL	F1
			Synesthetic			
SR-BiLSTM	43.7	54.2	63.8	64.9	0.0	54.8
E2EB-iLSTM	47.6	60.0	66.6	65.9	33.7	59.3
PF-BERT	70.4	67.2	76.4	81.6	49.5	73.2
MelBERT	66.8	71.3	74.9	78.1	24.3	70.7
MrBERT	62.0	67.2	74.1	81.9	48.4	70.3
Ours(character)	66.2	64.2	73.9	81.2	**56.9**	71.0
Ours(character+radical)	71.2	**76.3**	76.6	77.3	45.2	74.1
Ours(character+radical+word)	**74.0**	71.4	77.8	81.2	45.8	75.0
Ours(character+radical+word+pos tag)	71.5	72.0	**78.1**	**82.5**	50.5	**75.1**

tag") achieves the best performances in the detection for both original modalities and synesthetic modalities.

- It is hard for the models to predict hearing and smell, especially for the detection of original sensory modalities.
- Notably, Jiang *et al.* (2022)'s work reported higher performances in both original modality detection and synesthetic modality detection. The improvements can be attributed to their use of golden standard annotations for the initial phase of the sensory word extraction task, which involves identifying the boundaries of words that evoke sensory experiences. The approach inadvertently enhanced the perceived accuracy of results in both original modality detection and synesthetic modality detection. Our model is designed as a sequential process aimed at automatically detecting and interpreting sensory-related words, such as those pertaining to taste or smell. The first step is to determine the boundaries of these sensory words. Subsequently, we ascertain the specific sense they relate to. In line with best practices for pipeline models, our model should not have access to the correct answers for the initial step during its operation, as the accurate label of the first task should not influence the prediction of the second task.

The result that the performances of the models for the detection of synesthetic modalities are better than those for the detection of original modalities may be caused by several factors. Firstly, detecting the original sensory modality mainly relies on the semantics of the sensory word, where the radical information makes a great contribution to the proposed model. However, the synesthetic sensory modality can be inferred from both the sensory word and the context. As demonstrated in Table 3, an inconsistency in the sensory modalities in a sentence can suggest a linguistic synesthetic usage, where an adjective generally shows the source modality and a noun the target modality. In addition, there are directional patterns between the five sensory modalities for linguistic synesthesia, as shown in Figure 2. That is, the probability of each sensory modality being used for another sensory modality is different. Zhao (2020)' corpus-based study showed that 84.9 percent of tactile adjectives were used for vision, 76.2 percent of gustatory adjectives for smell, and 87.9 percent of visual adjectives for hearing, while a very limited number of auditory and olfactory adjectives were used for other sensory modalities. Thus, the target synesthetic modality can also be inferred from the original sensory modality. With respect to hearing and

Table 7. The results of our proposed model with the sub-set of testing data with respect to the original modality, where "(Num.)" means the number of data from one original modality to one synesthetic modality

	Synesthetic(Num.)					
Original	TOUCH	TASTE	VISION	HEARING	SMELL	F1
TOUCH	75.1(231)	66.7(15)	68.7(134)	85.3(74)	76.2(10)	74.6
TASTE	4.9(5)	78.6(207)	77.4(135)	84.7(38)	40.0(31)	75.0
VISION	70.5(118)	22.2(1)	81.9(264)	75.1(136)	44.4(9)	76.9
HEARING	–(0)	54.5(6)	–(0)	–(0)	57.1(2)	56.2
SMELL	–(0)	60.0(11)	–(0)	–(0)	53.3(11)	56.7
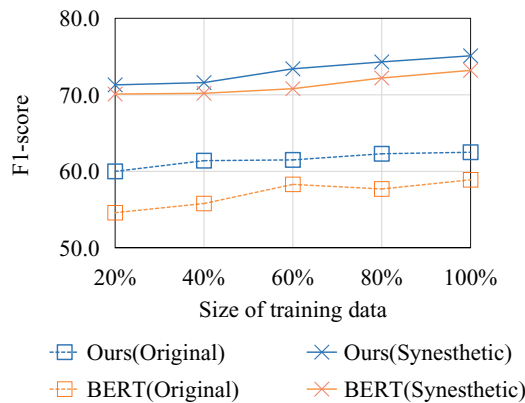
Figure 8. Influence of the size of the training data.

smell, as they are the most likely to act as the target domain in linguistic synesthesia, the performances of our proposed model for detecting hearing and smell as the synesthetic target modality are improved drastically, as shown in Table 6.

5.3.3 Analysis of factors

Table 7 presents an analysis of the number of data from one original modality to one synesthetic modality. In particular, the sensory modalities of hearing and smell have very limited numbers of linguistic synesthetic data. However, the sparse data on hearing and smell mirrors the crucial characteristic of human cognition that hearing and smell are hardly used as the target domains of linguistic synesthesia cross-linguistically (Strik Lievers 2015; Zhao 2020).

As sensory labeling and synesthesia labeling are usually expensive, we would like to test whether our model can still reach a reasonable performance with less data. Figure 8 shows the impact of the size of the training data on the performance of our model. Thus, our model is generally stable regardless of the size of the training data and the modality. This suggests that the linguistic features leveraged by this study contribute to a more robust model in linguistic synesthesia detection. In addition, the performances of both the BERT model and our proposed model increase with the size of the training data, which is in line with the general text classification models.

6. Conclusion

This study refines the NLP task called Chinese synesthesia detection proposed by Jiang *et al.* (2022). In particular, we construct a large-scale manually annotated Chinese synesthesia dataset, which will be released in the open resource platform of OSF. Based on the dataset, we incorporate culturally enriched linguistic features (i.e., character and radical information, word segmentation information, and POS tagging features) into a neural network model to detect linguistic synesthesia automatically. In terms of identifying the boundary of sensory words and jointly detecting the original and synesthetic sensory modalities of the words, our proposed model achieves state-of-the-art results on the dataset for linguistic synesthesia detection through extensive experiments. Furthermore, this study shows several linguistic features that are useful in the detection of linguistic synesthesia. That is, except for the radical information that is dependent on the Chinese writing system, the word segmentation and POS features could also be incorporated for the detection of linguistic synesthesia in other languages. Thus, our proposed model would be applied to the detection of linguistic synesthesia in other languages.

7. Limitations and future work

One of the limitations of this study is that our proposed model performs poorly with the subsets of the testing data for hearing and smell due to the data sparsity. Our future work will investigate how to integrate few-shot learning or data-augmenting methods in these two sparse data categories for linguistic synesthesia detection.

Secondly, the model, leveraging non-language-specific linguistic features, can detect linguistic synesthesia in various languages. However, evaluating it solely on Chinese may limit its generalizability due to its reliance on sensory-rich Chinese characters. This reliance could hinder its applicability across different languages, especially if key linguistic features are missing or inaccurately annotated. Future research should explore techniques like data augmentation or few-shot learning within a large language modeling framework to address limited annotated resources. This could enhance the model's versatility for application in diverse linguistic contexts.

Thirdly, our model is specifically designed for transformers based on the encoder structure and is not suitable for the encoder-decoder architecture of transformers, such as T5 (Raffel *et al.* 2020), GPT (Radford *et al.* 2018), etc. A potential limitation of our approach is that it is currently incompatible with these types of models, which may limit its applicability in scenarios where encoder-decoder models are preferred for their generative capabilities. Moving forward, a significant area for future work will be to explore strategies for integrating the model with the decoder component of these architectures and expanding the range of models that can benefit from the enrichment of sensory information.

References

Broadwell G. A., **Boz U.**, **Cases I.**, **Strzalkowski T.**, **Feldman L.**, **Taylor S.**, **Shaikh S.**, **Liu T.**, **Cho K. and Webb N.** (2013). Using imageability and topic chaining to locate metaphors in linguistic corpora. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, Springer, pp. 102–110.

Bulat L., **Clark S. and Shutova E.** (2017). Modelling metaphor with attribute-based semantics. In Short Papers *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Valencia, Spain: Association for Computational Linguistics, vol. 2, Short Papers, pp. 523–528, Short Papers

Cacciari C. (2008). Crossing the senses in metaphorical language. *The Cambridge Handbook of Metaphor and Thought* **55**(4), 425–443.

Chao Y. R. (1968). *A Grammar of Spoken Chinese*. Beijing: The Commercial Press.

Che W., **Li Z. and Liu T.** (2010). Ltp: a chinese language technology platform. In *Coling 2010: Demonstrations*, pp. 13–16.

Chen I., **Long Y.**, **Lu Q.**, **Huang C.-R.** and *et al.* (2021). Orthographic features for emotion classification in chinese in informal short texts. *Language Resources and Evaluation* **55**(2), 329–352.

Chen I.-H., **Long Y.**, **Lu Q. and Huang C.-R.** (2017). Leveraging eventive information for better metaphor detection and classification. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pp. 36–46.

Chen I.-H., **Zhao Q.**, **Long Y.**, **Lu Q. and Huang C.-R.** (2019). Mandarin chinese modality exclusivity norms. *PloS One* **14**(2), e0211336.

Chen K.-J., **Huang C.-R.**, **Chang L.-P. and Hsu H.-L.** (1996). Sinica corpus: Design methodology for balanced corpora. In *Proceedings of the 11th Pacific Asia Conference on Language, Information and Computation*, pp. 167–176.

Chen Z. and Wu K. (2020). ForceReader: A BERT-based interactive machine reading comprehension model with attention separation. In *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online), pp. 2676–2686, International Committee on Computational Linguistics.

Choi M., **Lee S.**, **Choi E.**, **Park H.**, **Lee J.**, **Lee D. and Lee J.** (2021). Melbert: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Chou Y.-M. and Huang C.-R. (2006). Hantology-a linguistic resource for Chinese language processing and studying. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*.

Chou Y.-M. and Huang C.-R. (2010). Hantology: Conceptual system discovery based on orthographic convention. In *Ontology and the Lexicon: A Natural Language Processing Perspective*, pp. 122–143.

Cytowic R. E. (2002). *Synesthesia: A Union of the Senses*. Cambridge: MIT Press.

Devlin J., **Chang M.-W.**, **Lee K. and Toutanova K.** (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In (Long and Short Papers)*Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota: Association for Computational Linguistics, vol. **1 (Long and Short Papers)**, pp. 4171–4186.

Dong Z. and Dong Q. (2003). Hownet-a hybrid language and knowledge resource. In *International Conference on Natural Language Processing and Knowledge Engineering*, Proceedings, 2003. IEEE, pp. 820–824.

Duan Y. (1815). *Commentary on Explaining Graphs and Analyzing Characters*. Nanjing: Phoenix Press.

Fleiss J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin* **76**(5), 378–382.

Gao G., **Choi E.**, **Choi Y. and Zettlemoyer L.** (2018). Neural metaphor detection in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium: Association for Computational Linguistics, pp. 607–613.

Gong H., **Gupta K.**, **Jain A. and Bhat S.** (2020). IlliniMet: Illinois system for metaphor detection with contextual and linguistic information. In *Proceedings of the Second Workshop on Figurative Language Processing*, Online. Association for Computational Linguistics, pp. 146–153.

Hercig T. and Lenc L. (2017). The impact of figurative language on sentiment analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, Varna, Bulgaria: INCOMA Ltd, pp. 301–308.

Hou R., **Huang C.-R. and Liu H.** (2019). A study on chinese register characteristics based on regression analysis and text clustering. *Corpus Linguistics and Linguistic Theory* **15**(1), 1–37.

Hu M., **Peng Y.**, **Huang Z.**, **Li D. and Lv Y.** (2019). Open-domain targeted sentiment analysis via span-based extraction and classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, pp. 537–546.

Huang C.-R. and Shi D. (2016). *A Reference Grammar of Chinese*. Cambridge: Cambridge University Press.

Huang C.-R., **Šimon P.**, **Hsieh S.-K. and Prévot L.** (2007). Rethinking Chinese word segmentation: Tokenization, character classification, or wordbreak identification. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, Prague, Czech Republic: Association for Computational Linguistics, pp. 69–72.

Jiang X., **Zhao Q.**, **Long Y. and Wang Z.** (2022). Chinese synesthesia detection: New dataset and models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 3877–3887.

Kingma D. P. and Ba J. (2015). Adam: A method for stochastic optimization. In *ICLR (Poster)*

Kumcu A. (2021). Linguistic synesthesia in Turkish: A corpus-based study of crossmodal directionality. *Metaphor and Symbol* **36**(4), 241–255.

Lafferty J. D., **McCallum A. and Pereira F. C. N.** (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc, vol **ICML '01**, pp. 282–289.

Lample G., **Ballesteros M.**, **Subramanian S.**, **Kawakami K. and Dyer C.** (2016). Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*, pp. 260–270.

Li S., **Zhou G. and Huang C.-R.** (2012). Active learning for Chinese word segmentation. In *Proceedings of COLING 2012: Posters*, Mumbai, India, pp. 683–692, The COLING. 2012 Organizing Committee

Li X., **Feng J.**, **Meng Y.**, **Han Q.**, **Wu F. and Li J.** (2020). A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics, pp. 5849–5859.

Liu H., **Lievers F. S. and Huang C.-R.** (2015). Automatic extraction and mapping directionality of synaesthetic sentences of modern Chinese. *Computer Engineering & Science* **37**(12), 2294–2299.

Mikolov T., **Chen K.**, **Corrado G. and Dean J.** (2013). Efficient estimation of word representations in vector space.

Miller G. and Johnson-Laird P. (1976). *Language and Perception*. Cambridge: Cambridge University Press.

Picard R. W. (2000). *Affective Computing*. Cambridge: MIT Press.

Popova Y. (2005). Image schemas and verbal synaesthesia. *From Perception to Meaning: Image Schemas in Cognitive Linguistics* **29**, 395–419.

Purves D., **Augustine G. J.**, **Fitzpatrick D.**, **Katz L. C.**, **LaMantia A.-S.**, **McNamara J. O. and Williams S. M.** (2001). *Neuroscience*. Massachusetts: Sinauer Associates, Inc.

Radford A., **Narasimhan K.**, **Salimans T.**, **Sutskever I.**, *et al.* (2018). Improving language understanding by generative pre-training.

Raffel C., **Shazeer N.**, **Roberts A.**, **Lee K.**, **Narang S.**, **Matena M.**, **Zhou Y.**, **Li W. and Liu P. J.** (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* **21**(1), 5485–5551.

Ramachandran V. S. and Hubbard E. M. (2001). Synaesthesia–a window into perception, thought and language. *Journal of Consciousness Studies* **8**(12), 3–34.

Rizzo M. (1989). Synesthesia: A union of the senses. *Neurology* **39**(10), 1413–1413.

Schuster M. and Paliwal K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions On Signal Processing* **45**(11), 2673–2681.

Shen Y. (1997). Cognitive constraints on poetic figures.

Shen Y. and Cohen M. (1998). How come silence is sweet but sweetness is not silent: a cognitive account of directionality in poetic synaesthesia. *Language and Literature* **7**(2), 123–140.

Shen Y. and Gil D. (2008). Sweet fragrances from Indonesia: A universal principle governing directionality in synaesthetic metaphors. In van Peer W and Auracher J. (eds), *New Beginnings in Literary Studies*, Newcastle, UK: Cambridge Scholars Publishing, pp. 49–71.

Simner J. and Hubbard E. (2013). *The Oxford Handbook of Synesthesia*. Oxford: Oxford University Press.

Song W., **Zhou S.**, **Fu R.**, **Liu T. and Liu L.** (2021). Verb metaphor detection via contextual relation learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online. Association for Computational Linguistics, pp. 4240–4251.

Strik Lievers F. (2015). Synaesthesia: A corpus-based study of cross-modal directionality. *Functions of Language* **22**(1), 69–95.

Strik Lievers F. (2017). Figures and the senses. *Review of Cognitive Linguistics* **15**(1), 83–101.

Strik Lievers F. and Huang C.-R. (2016). A lexicon of perception for the identification of synaesthetic metaphors in corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 2270–2277.

Strik Lievers F., **Xu G. and Xu H.** (2013). A methodology for the extraction of lexicalized synaesthesia from corpora. In *ICL 19 (International Congress of Linguists)*. Geneva, Switzerland, pp. 21–27.

Su C., **Wu K. and Chen Y.** (2021). Enhanced metaphor detection via incorporation of external knowledge based on linguistic theories. In *Findings of the Association for Computational*. Linguistics: ACL-IJCNLP, Association for Computational Linguistics, Online, pp. 1280–1287.

Tang D., **Qin B.**, **Feng X. and Liu T.** (2016). Effective LSTMs for target-dependent sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 3298–3307.

Tsvetkov Y., **Boytsov L.**, **Gershman A.**, **Nyberg E. and Dyer C.** (2014). Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Baltimore, Maryland: Association for Computational Linguistics, pp. 248–258.

Turner S. and Littlemore J. (2023). The many faces of creativity: exploring synaesthesia through a metaphorical lens. In *Elements in Cognitive Linguistics*. Cambridge University Press.

Turney P., **Neuman Y.**, **Assaf D. and Cohen Y.** (2011). Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland: UK. Association for Computational Linguistics, pp. 680–690.

Ullmann S. (1957). *The Principles of Semantics*. Oxford: Basil Blackwell.

Wan M., **Ahrens K.**, **Chersoni E.**, **Jiang M.**, **Su Q.**, **Xiang R. and Huang C.-R.** (2020). Using conceptual norms for metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pp. 104–109.

Wang H., **Gan Z.**, **Liu X.**, **Liu J.**, **Gao J. and Wang H.** (2019). Adversarial domain adaptation for machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China: Association for Computational Linguistics, pp. 2510–2520.

Wang Y., **Huang M.**, **Zhu X. and Zhao L.** (2016). Attention-based ISTM for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 606–615.

Weitzel L., **Prati R. C. and Aguiar R. F.** (2016). *The Comprehension of Figurative Language: What Is the Influence of Irony and Sarcasm on NLP Techniques?*. Cham: Springer International Publishing, pp. 49–74.

Williams J. M. (1976). Synaesthetic adjectives: A possible law of semantic change. *Language* **52**(2), 461–478.

Winter B. (2019a). *Sensory Linguistics: Language, Perception and Metaphor*. Amsterdam: John Benjamins Publishing Company.

Winter B. (2019b). Synaesthetic metaphors are neither synaesthetic nor metaphorical. *Perception Metaphors* **19**, 105–126.

Woon W. L. and Yun W. (1987). *Chinese Writing: Its Origin and Evolution*, vol. 2. Yamaguchi: University of East Asia Press.

Wu C., Wu F., Chen Y., Wu S., Yuan Z. and Huang Y. (2018). Neural metaphor detecting with CNN-LSTM model. In *Proceedings of the Workshop on Figurative Language Processing*, New Orleans, Louisiana: Association for Computational Linguistics, pp. 110–114.

Xiang R., Li J., Wan M., Gu J., Lu Q., Li W. and Huang C.-R. (2021). Affective awareness in neural sentiment analysis. *Knowledge-Based Systems* **226**, 107137.

Xu S. (156). *Explaining Graphs and Analyzing Characters*. Beijing: Zhonghua Book Company.

Xu T. (2005). *Hanyu Jiegou de Jiben Yuanli*. Shandong: China Ocean University Press.

Ye W. (2015). *Hanyu Jiegou de Jiben Yuanli*. Taiwan: Huamulan Culture Press.

Yu N. (2003). Synesthetic metaphor: A cognitive perspective.

Zhang S., Zhang X., Chan J. and Rosso P. (2019). Irony detection via sentiment-based transfer learning. *Information Processing & Management* **56**(5), 1633–1644.

Zhao Q. (2020). *Embodied Conceptualization or Neural Realization: A Corpus-Driven Study of Mandarin Synaesthetic Adjectives*. Singapore: Springer.

Zhao Q., Ahrens K. and Huang C.-R. (2022). Linguistic synesthesia is metaphorical: a lexical-conceptual account. *Cognitive Linguistics* **33**(3), 553–583.

Zhao Q., Huang C.-R. and Ahrens K. (2019). Directionality of linguistic synesthesia in mandarin: A corpus-based study. *Lingua* **232**, 102744.

Zhao Q., Huang C.-R. and Long Y. (2018). Synaesthesia in Chinese: A corpus-based study on gustatory adjectives in mandarin. *Linguistics* **56**(5), 1167–1194.

Zhong Y., Wan M., Ahrens K. and Huang C.-R. (2022). Sensorimotor norms for chinese nouns and their relationship with orthographic and semantic variables. *Language, Cognition and Neuroscience* **37**(8), 1–23.

Zolyomi A. and Snyder J. (2021). Social-emotional-sensory design map for affective computing informed by neurodivergent experiences. *Proceedings of the ACM on Human-Computer Interaction* **5**(CSCW1), 1–37.