Generating Bilingual Pragmatic Color References

Will Monroe
Computer Science Department
Stanford University
wmonroe4@cs.stanford.edu

Jennifer Hu
Department of Mathematics
Harvard University
jenniferhu@college.harvard.edu

Andrew Jong
Department of Computer Science
San Jose State University
andrewjong.cs@gmail.com

Christopher Potts
Department of Linguistics
Stanford University
cgpotts@stanford.edu

Abstract

Contextual influences on language exhibit substantial language-independent regularities; for example, we are more verbose in situations that require finer distinctions. However, these regularities are sometimes obscured by semantic and syntactic differences. Using a newly-collected dataset of color reference games in Mandarin Chinese (which we release to the public), we confirm that a variety of constructions display the same sensitivity to contextual difficulty in Chinese and English. We then show that a neural speaker agent trained on bilingual data with a simple multitask learning approach displays more human-like patterns of context dependence and is more pragmatically informative than its monolingual Chinese counterpart. Moreover, this is not at the expense of language-specific semantic understanding: the resulting speaker model learns the different basic color term systems of English and Chinese (with noteworthy crosslingual influences), and it can identify synonyms between the two languages using vector analogy operations on its output layer, despite having no exposure to parallel data.

1 Introduction

In grounded communication tasks, speakers face pressures in choosing referential expressions that distinguish their targets from others in the context, leading to many kinds of pragmatic meaning enrichment. For example, the harder a target is to identify, the more the speaker will feel the need to refer implicitly and explicitly to alternatives to draw subtle contrasts (Zipf, 1949; Horn, 1984; Levinson, 2000). However, the ways in which these contrasts are expressed depend heavily on language-specific syntax and semantics.
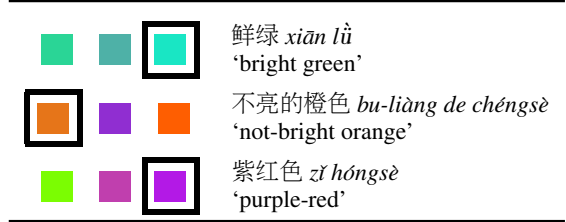

Figure 1: Reference game contexts and utterances from our Chinese corpus. The boxed color is the target. Some color terms show differences between Chinese and English, such as 绿 *lǜ* 'green' in the first example for a color that might be referred to with 'blue' or 'aqua' in English.

In this paper, we seek to develop a model of contextual language production that captures language-specific syntax and semantics while also exhibiting language-independent patterns of responsiveness to contextual differences. We focus on a color reference game (Rosenberg and Cohen, 1964; Dale and Reiter, 1995; Krahmer and van Deemter, 2012) played in both English and Mandarin Chinese. A reference game (Figure 1) involves two agents, one designated the "speaker" and the other the "listener". The speaker and listener are shown the same set of k colors $C = \{c_1, \ldots, c_k\}$ (in our experiments, $k = 3$), and one of these colors c_t is indicated secretly to the speaker as the "target". Both players share the same goal: that the listener correctly guesses the target color. The speaker may communicate with the listener in free-form natural-language dialogue to achieve this goal. Thus, a model of the speaker must process representations of the colors in the context and produce an utterance to distinguish the target color from the others. We evaluate a

sequence-to-sequence speaker agent based on that of Monroe et al. (2017), who also collected the English data we use; our Chinese data are new and were collected according to the same protocols.

English and Chinese exhibit substantial commonalities but also important differences. While both use fairly similar syntax for color descriptions, our reference game is designed to elicit constructions that make reference to the context, and these constructions—particularly comparatives and negation—differ morpho-syntactically and pragmatically between the two languages. Additionally, Chinese is considered to have a smaller number of basic color terms (Berlin and Kay, 1969), which predicts markedness of more specific descriptions.

Our primary goal is to examine the effects of *bilingual* training: building one speaker trained on both English and Chinese data with a shared vocabulary, so that it can produce utterances in either language. The reference game setting offers an objective measure of success on the grounded language task, namely, the speaker's ability to guide the listener to the target. We use this to address the tricky problem of speaker evaluation. Specifically, we use the speaker model and an application of Bayes' rule to infer the most likely target color given a human utterance, and we report the accuracy of that process at identifying the target color. We refer to this metric as *pragmatic informativeness* because it requires not only accuracy but also effectiveness at meeting the players' shared goal (Grice, 1975). A more formal definition and a discussion of alternatives are given in Section 4.1.

We show that a bilingually-trained model produces distributions over Chinese utterances that have higher pragmatic informativeness than a monolingual model, suggesting that it has leveraged the bilingual data to learn relevant abstract pragmatic abilities. The same model learns color synonyms between the two languages without being directly exposed to labeled pairs. Using a color term elicitation task from Berlin and Kay (1969) on our models, we show that the learned lexical meanings are largely faithful to each language's basic color system but show subtle influences from the other language, reinforcing the claim that the model incorporates both language-specific semantics and language-general pragmatics.

2 Data collection

We adapted the open-source reference game framework of Hawkins (2015) to Chinese and followed the data collection protocols of Monroe et al. (2017) as closely as possible, in the hope that this can be the first step in a broader multilingual color reference project. We recruit pairs of players on Amazon Mechanical Turk in real time, randomly assigning one the role of the speaker and the other the listener. Players are self-reported Chinese speakers, but they must pass a series of Chinese comprehension questions in order to proceed, with instructions in a format preventing copy-and-paste translation. The speaker and listener are placed in a game environment in which they both see the three colors of the context and a chatbox. The speaker sends messages through the chatbox to describe the target to the listener, who then attempts to click on the target. This ends the round, and three new colors are generated for the next. Both players can send messages through the chatbox at any time. After filtering out extremely long messages (number of tokens greater than 4σ above the mean), spam games,[1] and players who self-reported confusion about the game, we have a new corpus of 5,774 Chinese messages in color reference games, which we will release publicly. Data management information is given in Appendix B.

As in Monroe et al. (2017), the contexts are divided into three groups of roughly equal size: in the *far* condition (1,421 contexts), all the colors are at least a threshold distance θ from each other; in the *split* condition (1,412 contexts), the target and one distractor are less than θ from each other, with the other distractor at least θ away from both; and in the *close* condition (1,425 contexts), all colors are within θ from each other. We set $\theta = 20$ by the CIEDE2000 color-difference formula (Sharma et al., 2005), with all colors different by at least 5.

3 Human data analysis

As we mentioned earlier, our main goal with this work is to investigate the effects of bilingual training on pragmatic language use. We first examine the similarities and differences in pragmatic behaviors between the English and Chinese corpora we use. The picture that emerges accords well

---

[1]Some players found they could advance through rounds by sending duplicate messages. Games were considered spam if the game contained 25 or more duplicates.
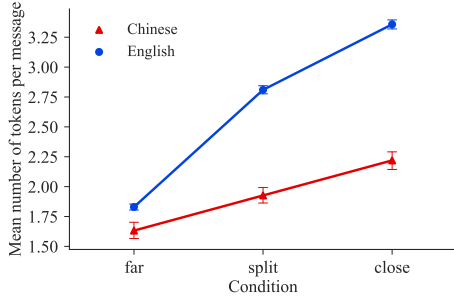
Figure 2: Comparison of mean length of messages in English and Chinese. The *split* and *close* conditions have more similar context colors (Section 2).

Figure 3: Comparison of WordNet specificity in Chinese and English.

with our expectations about pragmatics: the broad patterns are aligned across the two languages, with the observed differences mostly tracing to the details of their lexicons and constructions.

3.1 Message length

We expect message length to correlate with the difficulty of the context: as the target becomes harder to distinguish from the distractors, the speaker will produce more complex messages, and length is a rough indicator of such complexity. To test this hypothesis, we used the Natural Language Toolkit (NLTK; Bird et al. 2009) and Jieba (Junyi, 2015) to tokenize English and Chinese messages, respectively, and counted the number of tokens in both languages as a measure of message length. The results (Figure 2) confirm that in both languages, players become more verbose in more difficult conditions.[2]

3.2 Specificity

In the *split* and *far* conditions, the speaker must make fine-grained distinctions. A broad color term like *red* will not suffice if there are two reds, but more specific terms like *maroon* might identify the target. Thus, we expect specificity to increase as the difficulty of the context does. To assess this, we use WordNet (Fellbaum, 1998) to transform adjectives into derivationally-related noun forms, filter for nouns with *color* in their hypernym paths, and mark a message as "specific" if it contains at least one word with a hypernym depth greater than 7. For Chinese, we translate to English via Google Translate, then measure the translated word using

WordNet. (Though Mandarin variations of Word-Net exist, we chose this translation method to standardize hypernym paths for both languages.)

Figure 3 summarizes the results of this measurement. In general, the usage of high-specificity color words increases in more difficult conditions, as expected. However, we see that Chinese speakers use them significantly less than English speakers. Instead, Chinese speakers use descriptive color words, such as 草 *cǎo* 'grass' and 海 *hǎi* 'ocean', which do not contain "color" in their hypernym paths and are thus not marked as high-specificity. To quantify this observation, we annotated random samples of 250 messages from each language for whether they contained descriptive color terms, and found that 0.4% of the English messages contain descriptive color words versus 15.26% of the Chinese messages.

We also observe that Chinese speakers use long strings of low-specificity color terms such as 土橘黄色 *tǔ jú huáng sè* 'earth tangerine yellow' to describe the target. This difference is arguably expected given the claims of Berlin and Kay (1969) and others that Chinese has fewer basic color terms than English. However, we wish to add a pragmatic twist to those generalizations: even where Chinese has an appropriate specific color term, it might be unusual enough that speakers avoid it: the term could confuse a listener, either by mere unfamiliarity or by implying that the choice of that specific, unusual color term was meant to emphasize a specific, unusual aspect of the color it refers to (markedness implicature).

3.3 Comparatives, superlatives, and negation

To detect comparative and superlative adjectives in English, we use NLTK POS-tagging, which outputs JJR and RBR for comparatives, and JJS and RBS for superlatives. In Chinese, we look for the

---

[2]We do not believe that the overall drop in message length from English to Chinese reflects a fundamental difference between the languages; this has a few possible explanations, from Chinese messages taking the form of "sentence segments" (Wang and Qin, 2010) to differences in tokenization.
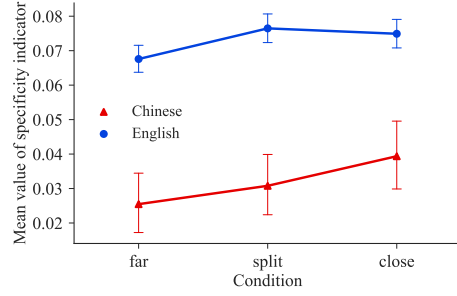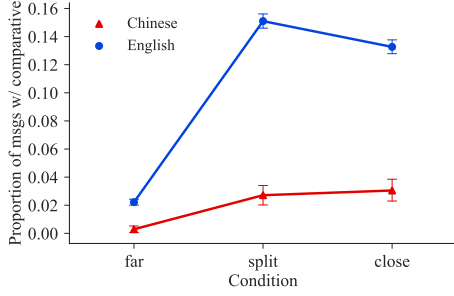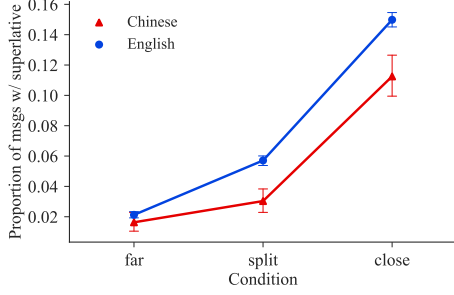
(a) Usage of comparative adjectives in Chinese and English.

(b) Usage of superlative adjectives in Chinese and English.
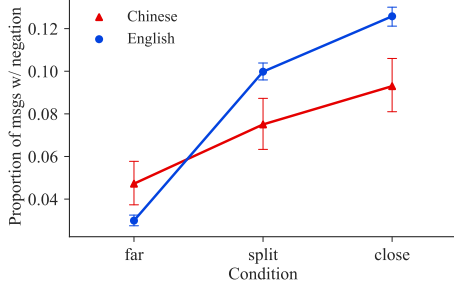
(c) Usage of negation in Chinese and English.

Figure 4: Comparison of usage of comparatives, superlatives, and negation in English and Chinese.

tokens 更 *gèng* 'more' and 比 *bǐ* 'comparatively' to detect comparatives and 最 *zuì* 'most' to detect superlatives. We detect negation by tokenizing messages with NLTK and Jieba and then looking for the tokens *not* and *n't* in English and corresponding 不 *bù* and 没 *méi* in Chinese.

Both languages exhibit similar trends for superlative adjectives. In English, comparatives are used most frequently in the *split* condition and second most frequently in the *close* condition, while in Chinese, they occur at around the same rate in the *split* and *close* conditions. The literature is not conclusive about the source of these differences. Xia (2014) argues that complex attributives are rarely used and sound "syntactically deviant or Europeanized" (Zhu, 1982; Xie, 2001) in Chinese, citing the left-branching nature of the language as restricting attributives in length and

complexity. There are also conflicting theories on the markedness of gradable adjectives in Chinese (Grano, 2012; Ito, 2008); such markedness may contribute to the frequency at which comparative forms are used.

We also see that both languages follow the same general trend of using negation more frequently as the condition becomes more difficult.

4 Models

We build and evaluate three artificial agents on this reference game task, two trained on monolingual descriptions (one for each language) and one on bilingual descriptions. We base these models on the basic speaker architecture from Monroe et al. (2017). The monolingual speakers represent the context by passing all the context colors as input to a long short-term memory (LSTM) sequence encoder, then concatenating this representation with a word vector for each previous output token as the input to an LSTM decoder that produces a color description token-by-token. This defines a distribution over descriptions u conditioned on the target and context, $S(u \mid c_t, C)$.

To accommodate bilingual training with this architecture, we expand the vocabulary to include English and Chinese words, and we add a flag ℓ to the input specifying whether the model's output should be in English ($\ell = 0$) or Chinese ($\ell = 1$):

$$S(u \mid \ell, c_t, C) = \prod_{i=1}^{|u|} s(u_i \mid u_{1..i-1}, \ell, c_t, C)$$

The flag ℓ is embedded as a single additional dimension that is concatenated alongside the context and input (previous token) vectors for the encoder. See Appendix A for additional training details.

4.1 Pragmatic informativeness

As mentioned in Section 1, we evaluate the two models on a measure of *pragmatic informativeness*: how well does the model represent a human speaker, such that a generative model of a listener can be built from it to interpret utterances? Formally, for a speaker $S(u \mid \ell, c_t, C)$ and an example consisting of an utterance, language identifier, and color context (u, ℓ, C), we identify the t^* that maximizes the probability of u according to S:

$$t^* = \arg\max_t S(u \mid c_t, C)$$

That is, L uses a noisy-channel model with a uniform prior over target colors and S as a generation

model to infer the most likely target color given the input utterance. The pragmatic informativeness of a speaker is the proportion of target colors in a test set correctly identified by t^*.

One drawback of this metric is it does not evaluate how faithful the model is to the overall distribution of human utterances, only the relative conditional likelihoods of human utterances for different target colors. In practice, since the agents are trained to minimize log likelihood, we do not observe our agents frequently producing wildly unhumanlike utterances; however, this is a caveat to keep in mind for evaluating agents that do not naturally approximate a language model.

The understanding model implied in this metric is equivalent to a version of the Rational Speech Acts model of pragmatic language understanding (Frank and Goodman, 2012; Goodman and Frank, 2016), applied to the output of a neural network model. A related metric is *communicative success* as defined by Golland et al. (2010), which judges the speaker by the accuracy of a human listener when given model-produced utterances. Our pragmatic informativeness metric instead gives a model-derived listener human utterances and assesses its accuracy at identifying colors. Pragmatic informativeness has the advantage of not requiring additional expensive human labeling in response to model outputs; it can be assessed on an existing collection of human utterances, and can therefore be considered an automatic metric.

4.2 A note on perplexity and token metrics

Perplexity is a common intrinsic evaluation metric for generation models. However, for comparing monolingual and bilingual models, we found perplexity to be unhelpful, owing largely to its vocabulary-dependent definition. Specifically, if we fix the vocabulary in advance to include tokens from both languages, then the monolingual model performs unreasonably poorly, and bilingual training helps immensely. However, this is an unfair comparison: the monolingual model's high perplexity is dominated by low probabilities assigned to rare tokens in the opposite-language data that it did not see. Thus, perplexity ceases to be a measure of language modeling ability and assumes the role of a proxy for the out-of-vocabulary rate.

On the other hand, if we define the output vocabulary to be the set of tokens seen at least n times in training ($n = 1$ and 2 are common), then

test	train	dev acc	test acc
en	en	**80.51**	**83.06**
	en+zh	79.73	81.43
zh	zh	67.16	67.75
	en+zh	**71.81**	**72.89**

Table 1: Pragmatic informativeness scores (%) for monolingual and bilingual speakers.

monolingual training yields better perplexity than bilingual training, but mainly because including opposite-language training data forces the bilingual model to predict more rare words that would otherwise be replaced with ⟨unk⟩.[3] This produces the counterintuitive result that perplexity initially goes *up* (gets worse) when increasing the amount of training data. (As a pathological case, with no training data, a model can get a perfect perplexity of 1 by predicting ⟨unk⟩ for every token.)

Another alternative for evaluating speaker models is token overlap metrics such as word error rate (WER) and BLEU (Papineni et al., 2002). These were at or worse than chance despite qualitatively adequate speaker outputs, due to the high diversity in valid speaker outputs for similar contexts. This problem is common in dialogue tasks, for which BLEU is known to be an ineffective speaker evaluation metric (Liu et al., 2016).

5 Experimental results and analysis

Pragmatic informativeness of the models on English and Chinese data is shown in Table 1. The main result is that training a bilingual model helps compared to a Chinese monolingual one; however, the benefit is asymmetrical, as training on monolingual English data is superior for English data to training on a mix of Chinese and English. All differences in Table 1 are significant at $p < 0.001$ (approximate permutation test, 10,000 samples; Padó, 2006), except for the decrease on the English dev set, which is significant at $p < 0.05$.

An important difference between our corpora is that the English dataset is an order of magnitude larger than the Chinese. Intuitively, we expect adding more training data on the same task will improve the model, regardless of language.

---

[3]The rare words that make this difference are primarily the small number of English words that were used by the Chinese-language participants; no Chinese words were observed in the English data from Monroe et al. (2017)
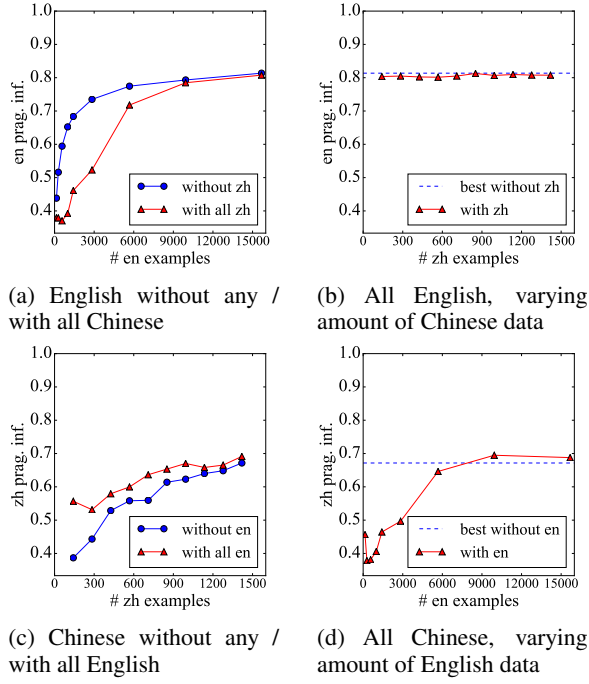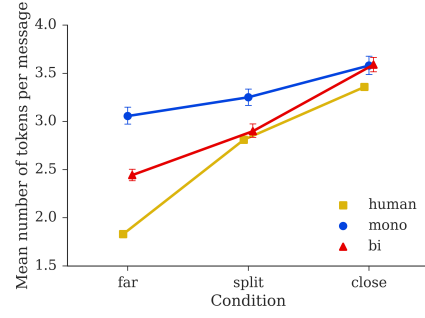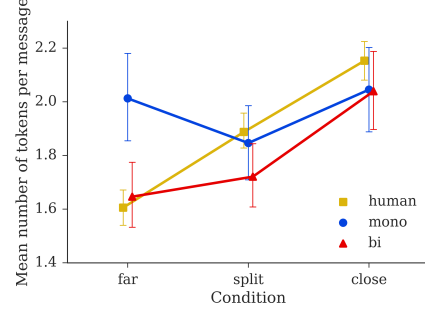
(a) English without any / with all Chinese

(b) All English, varying amount of Chinese data

(c) Chinese without any / with all English

(d) All Chinese, varying amount of English data

Figure 5: Pragmatic informativeness (dev set) for different amounts and languages of training data.

(a) Human and model utterance lengths in English.

(b) Human and model utterance lengths in Chinese.

Figure 6: Comparison of mean length of messages between human and model utterances.

However, we find that the effect of dataset size is not so straightforward. In fact, the differences in training set size convey an asymmetric benefit. Figure 5 shows the pragmatic informativeness of the monolingual and bilingual speakers on the development set as a function of dataset size (number of English and Chinese utterances). The blue curves (circles) in the plots on the left, Figure 5a and Figure 5c, are standard learning curves for the monolingual models, and their parallel red curves (triangles) show the pragmatic informativeness of the bilingual model with the same amount of in-language data plus all available data in the opposite language. The plots on the right, Figure 5b and Figure 5d, show the effect of gradually adding opposite-language data to the bilingual model starting with all of the in-language data.

Overall, we see that adding all English data consistently helps the Chinese monolingual model, whereas adding all Chinese data consistently hurts the English monolingual model (though with diminishing effects as the amount of English data increases). Adding small amounts of English data—especially amounts comparable to the size of the Chinese dataset—decreases accuracy of the Chinese model dramatically. This suggests an interaction between the total amount of data and the effect of bilingual training: a model trained on

a moderately small number of in-language examples can benefit from a much larger training set in another language, but combining data in two languages is detrimental when both datasets are very small and has very little effect when the in-language training set is large. This implies a benefit primarily in low-resource settings, which agrees with the findings of Johnson et al. (2016) using a similar architecture for machine translation.

5.1 Comparing model and human utterances

Our hypothesis is that the bilingually-trained model is learning language-independent aspects of pragmatic language production. Figure 6 reveals that the bilingually-trained model better captures the pattern of increasing message length in harder conditions. In both languages, the monolingual model uses longer utterances in the easy *far* condition than human speakers do, whereas the bilingual model is significantly closer on that condition to the human statistics. We see similar results in the use of negations and comparatives; the use of superlatives is not substantially different between the monolingual and bilingual models.

zh	en	en	zh
绿色 'green'	**green**	green	绿 **'green'**
紫色 'purple'	**purple**	blue	蓝 **'blue'**
蓝色 'blue'	purple	purple	蓝 **'blue'**
灰色 'grey'	**grey**	bright	鲜艳 **'bright'**
亮 'bright'	**bright**	pink	粉色 **'pink'**
灰 'grey'	-er	grey	灰 **'grey'**
蓝 'blue'	*teal*	dark	暗 **'dark'**
绿 'green'	**green**	gray	灰 **'grey'**
紫 'purple'	**purple**	yellow	黄色 **'yellow'**
草 'grass'	*green*	light	最 'most'

Table 2: Bilingual lexicon induction from Chinese to English (first two columns) and vice versa (last two). Correct translations in **bold**, semantically close words in *italic*.

5.2 Bilingual lexicon induction

To get a better understanding of the influence of the bilingual training on the model's lexical representations in the two languages, we extracted the weights of the final softmax layer of the bilingual speaker model and used them to induce a bilingual lexicon with a word vector analogy task. For two pairs of lexical translations, 蓝色 *lánsè* → "blue" and "red" → 红 *hóng*, we took the difference between the source language word vector and the target language word vector. To "translate" a word, we added this "translation vector" to the word vector for the source word, and found the word in the opposite language with the largest inner product to the resulting vector. The results are presented in Table 2. We identified the 10 most frequent color-related words in each language to translate. (In other words, we did not use this process to find translations of function words like "the" or the Chinese nominalization/genitive particle 的 *de*, but we show proposed translations that were not color-related, such as 灰 *huī* being translated as the English comparative ending "-er".) The majority of common color words are translated correctly by this simple method, showing that the vectors in the softmax layer do express a linear correspondence between the representation of synonyms in the two languages.

5.3 Color term semantics

The above experiment suggests that the bilingual model has learned word semantics in ways that discover translation pairs. However, we wish to know whether bilingual training has resulted in changes to the model's output distribution reflecting differences in the two languages' color sys-

(a) Monolingual Chinese

(b) Chinese with English data
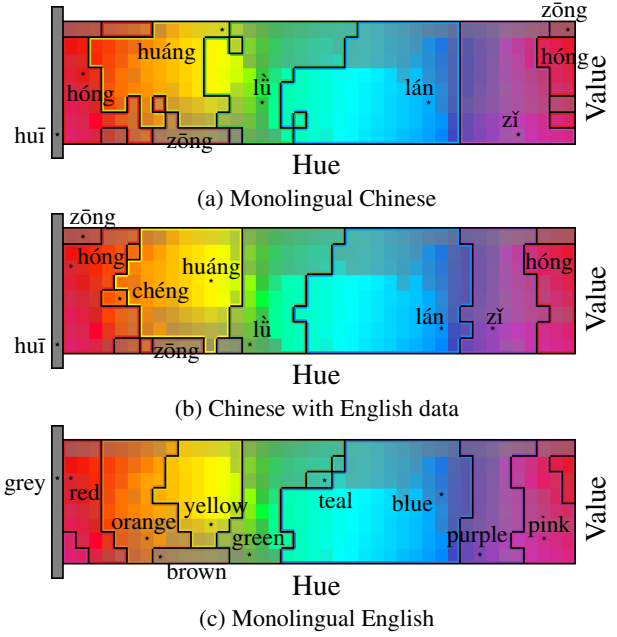
(c) Monolingual English

Figure 7: Color term lexica: colors in the World Color Survey palette grouped by highest-probability description, averaged over 10 randomly-generated pairs of distractor colors. The color that results in the highest probability of each description is marked with a star. English influences on the bilingual model include the appearance of 橙色 *chéngsè* 'orange' and narrowing of 黄色 *huángsè* 'yellow' and 绿色 *lǜsè* 'green'.

tems. To evaluate this, we performed an experiment similar to the basic color term elicitations in the World Color Survey (WCS; Berlin and Kay, 1969) on our models. For each of the 330 colors in the original WCS, we presented that color to our monolingual and bilingual models and recorded the most likely color description according to the conditional language model. Our models require a three-color context to produce a description; as an approximation to eliciting context-insensitive color terms, we gave the model ten contexts with randomly generated (uniform in H, S, and V) distractor colors and averaged the language model probabilities. We also identified, for each color term produced as the most likely description of one or more colors, the color that resulted in the highest probability of producing that term.

The results are in Figure 7. The charts use the layout of the WCS stimulus, in which the two axes represent dimensions of color variation similar to hue and lightness. Each region represents a set of colors that the model labeled with the same color term, and a star marks the color that resulted in the

highest probability of producing that term. The Chinese terms, except for 红 *hóng*, are abbreviated by deleting the final morpheme 色 *sè* 'color'.

The charts agree with Berlin and Kay (1969) on most of the differences between the two languages: *orange* and *pink* have clear regions of dominance in English, whereas in the Mandarin monolingual model *pink* is subsumed by 红 *hóng* 'red', and *orange* is subsumed by 黄色 *huángsè* 'yellow'. Our models produce three colors not in the six-color system[4] identified by Berlin and Kay for Mandarin: 灰色 *huīsè* 'grey', 紫色 *zǐsè* 'purple', and 棕色 *zōngsè* 'brown'. Of these, they explicitly exclude 灰色 *huīsè* from the set of basic color terms because it has a meaning that refers to an object ('ashes'); the other two may have been excluded for the same reason, or they may represent a change in the language or the influence of English on the participants' usage.[5]

A few differences between the monolingual and bilingual models can be characterized as an influence of one language's color system on the other. First, *teal* appears as a common description of a few color swatches from the English monolingual model, but the bilingual model, like the Chinese model, does not feature a common word for teal. Second, the Chinese monolingual model does not include a common word for orange, but the bilingual model identifies 橙色 *chéngsè* 'orange'. Finally, the English *green* is semantically narrower than the Chinese 绿色 *lǜsè*, and the Chinese bilingual model exhibits a corresponding narrowing of the range of 绿色 *lǜsè*. Changes to the lexical semantics due to opposite-language data should predict, all else equal, that the bilingual model would be a less effective model of a speaker in each language. The fact that we observe otherwise for the Chinese speaker suggests that the benefits of learning general pragmatic behaviors outweigh the drawbacks of cross-lingual semantic interference.

6   Related work

The method we use to build a bilingual model involves adding a single language dimension to the previous-token vectors in the encoder (Section 4).

Adding a hard constraint on the output vocabulary would make this equivalent to a simple form of multitask learning (Caruana, 1997; Collobert and Weston, 2008). (Our model occasionally ignores the flag and "code-switches" between the two languages within a single output, which is not possible in typical multitask architectures.)

Using shared parameters for cross-lingual representation transfer has a large literature. Klementiev et al. (2012) and Hermann and Blunsom (2014) use multitask learning with multilingual document classification to build cross-lingual word vectors, and observe accurate lexical translations from linear vector analogy operations. They include predicting translations for words in parallel data as one of their tasks. Our translations from vector relationships (Section 5.2) derive their cross-lingual relationships from the non-linguistic input of our grounded task, without parallel data.

Huang et al. (2013) note gains in speech recognition from cross-lingual learning with shared parameters. In machine translation, Johnson et al. (2016) add the approach of setting the output language using a symbol in the input. Kaiser et al. (2017) extend this to image captioning, speech recognition, and parsing in one multitask system. Our work complements these efforts with an indepth analysis of bilingual training on a grounded generation task and an exploration of the relationship between cross-lingual semantic differences and language-general pragmatics.

7   Conclusion

In this paper, we studied the effects of training on bilingual data in a grounded language task. We show evidence that bilingual training can be helpful, but with a non-obvious effect of dataset size: accuracy as a function of opposite-language data follows a U-shaped curve. The resulting model is more human-like in measures of language-general pragmatic effects, while exhibiting language-specific evidence of its bilingual training in the form of vector relationships between lexical pairs and influences of the opposite language on common color-term probabilities.

It should be noted that color descriptions in English and Chinese are similar both in their syntax and in the way they divide up the semantic space. We might expect that for languages like Arabic and Spanish (with their different placement of modifiers), or Waorani and Pirahã (with

---

[4]Notably absent are 'black' and 'white'. The collection methodology of Monroe et al. (2017) restricted colors to a single lightness, so black and white are not in the data. For these charts, we replaced the World Color Survey swatches with the closest color used in our data collection.

[5]MTurk's restriction to US workers makes English influence more likely than would otherwise be expected.

their much smaller color term inventories), the introduction of English data could have detrimental effects that outweigh the language-general gains. An investigation across a broader range of languages is desirable.

Our contribution includes a new dataset of human utterances in a color reference game in Mandarin Chinese, which we release to the public[6] with our code and trained model parameters.[7]

Acknowledgments

References

Brent Berlin and Paul Kay. 1969. *Basic Color Terms: their Universality and Evolution*. University of California Press, Berkeley and Los Angeles.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Sebastopol, CA.

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28:41–75.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 160–167. ACM.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Database*. MIT Press, Cambridge, MA.

Michael C. Frank and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998.

Dave Golland, Percy Liang, and Dan Klein. 2010. A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 Conference on Empirical Methods on Natural Language Processing (EMNLP)*.

Noah D. Goodman and Michael C. Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829.

Thomas Grano. 2012. Mandarin 'hen' and universal markedness in gradable adjectives. *Natural Language & Linguistic Theory*, 30(2):513–565.

H. Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry Morgan, editors, *Syntax and Semantics*, volume 3: Speech Acts, pages 43–58. Academic Press, New York.

Robert X. D. Hawkins. 2015. Conducting real-time multiplayer experiments on the web. *Behavior Research Methods*, 47(4):966–976.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 58–68, Baltimore, Maryland. Association for Computational Linguistics.

Laurence R. Horn. 1984. Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In Deborah Schiffrin, editor, *Meaning, Form, and Use in Context: Linguistic Applications*, pages 11–42. Georgetown University Press, Washington, D.C.

Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong. 2013. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7304–7308. IEEE.

Satomi Ito. 2008. Typology of comparatives. In *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation, PACLIC 22, Cebu City, Philippines, November 20-22, 2008*, pages 197–206.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2016. Google's multilingual neural machine translation system: enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*.

Sun Junyi. 2015. Jieba Python Library.

Łukasz Kaiser, Aidan N Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. 2017. One model to learn them all. *arXiv preprint arXiv:1706.05137*.

---

[6] https://cocolab.stanford.edu/datasets/colors.html
[7] https://github.com/futurulus/colors-in-context

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*, pages 1459–1474, Mumbai. International Committee on Computational Linguistics.

Emiel Krahmer and Kees van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Stephen C. Levinson. 2000. *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. MIT Press, Cambridge, MA.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP*.

Will Monroe, Robert X.D. Hawkins, Noah D. Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *TACL*, 5:325–338.

Sebastian Padó. 2006. *User's guide to `sigf`: Significance testing by approximate randomisation*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *ACL*.

Seymour Rosenberg and Bertram D. Cohen. 1964. Speakers' and listeners' processes in a word communication task. *Science*, 145(3637):1201–1203.

Gaurav Sharma, Wencheng Wu, and Edul N. Dalal. 2005. The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research & Application*, 30(1):21–30.

Kefei Wang and Hongwu Qin. 2010. A parallel corpus-based study of translational chinese. In Richard Xiao, editor, *Using Corpora in Contrastive and Translation Studies*, pages 164–181. Cambridge Scholars Publishing.

Yun Xia. 2014. *Normalization in Translation: Corpus-based Diachronic Research into Twentieth-century English Chinese Fictional Translation*. Cambridge Scholars Publishing.

Yaoji Xie. 2001. Hanyu yufa ouhua zongshu (A review of Europeanized Chinese grammar). *Yuwen yanjiu (Chinese Language Research)*, 1:17–22.

Dexi Zhu. 1982. *Yufa Jiangyi (Lectures on Grammar)*. The Commercial Press, Beijing.

George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge.

A Model details

Both Chinese monolingual and bilingual model were tuned for perplexity on a held-out subset of the training set, by random search followed by a local search from the best candidate until no single parameter change produced a better result. However, the tuned settings for the Chinese monolingual model did not outperform the settings from Monroe et al. (2017) for the English model on the development set, so in our final experiments the monolingual models used the same parameters.

The vocabulary for each model consisted of all tokens that were seen at least twice in training; the bilingual model's vocabulary is larger than the union of the words in each monolingual model because some tokens occurred once in each language (largely meta-commentary—e.g., *dunno*, *HIT*, *xD*—and some English color word typos).

hyperparameter	mono.	biling.
optimizer	ADAM	RMSProp
learning rate	0.004	0.004
dropout	0.1	0.1
gradient clip norm	–	1
LSTM cell size	100	50
embedding size	100	100
initial forget bias	0	5
nonlinearity	tanh	sigmoid
vocabulary size	895 (en) 260 (zh)	1,326

Table 3: Values of hyperparameters optimized in tuning for the monolingual and bilingual models, plus vocabulary sizes.

B Data management

This work was done under Stanford IRB Protocol 17827, which has the title "Pragmatic enrichment and contextual inference". Data was only collected from workers who indicated their informed consent. Workers on Amazon Mechanical Turk were paid $2.00 to complete each game consisting of 50 dialogue contexts, plus a bonus of $0.01 for each target the listener correctly identified. All worker identifiers have been removed from data that will be released; the only other information collected about the workers was their Chinese language proficiency.