# Topic marking in a Shanghainese corpus: from observation to prediction

WEIFENG HAN, ANTTI ARPPE and JOHN NEWMAN

*Abstract*
*Shanghainese is an extremely topic-prominent language with many topic markers in competition with one another, often without any obvious basis for the selection of one topic marker over another. We explore the influence of five variables on the five most frequent topic markers in a corpus of (spoken) Shanghainese: topic length, syntactic category of the topic, function of the topic, comment type, and genre. We carry out a multivariate statistical analysis of the data, relying on a polytomous logistic regression model. Our approach leads to a satisfying quantification of the role of each factor, as well as an estimate of the probabilities of combinations of factors, in influencing the choice of topic marker. This study serves simultaneously as an introduction to the* `polytomous` *package (Arppe 2013) in the statistical software package R.*

*Keywords: Shanghainese; Chinese dialect; topic marking; polytomous logistic regression; probabilities; statistical analysis*

## 1. Introduction

Languages subsumed under the general rubric of "Chinese", i.e., Chinese dialects as well as Mandarin (referring to the national language or *putonghua*), are well known for their preference for topic-comment structures in the syntactic organization of clause structure. As one might expect, though, the exact nature of topic marking can vary from one language to another, even within the Chinese language family. One finds, for example, variation with respect to the range of functions associated with topic marking (the formal element demarcating the topic at the right edge), as well as the number of topic markers available. In the context of variation in topic marking in Chinese, Shanghainese presents a particularly interesting example on account of the abundance of its topic markers, with more than twenty distinct topic markers available (Xu and Liu 2007 identify nine as being the most common). The basis for selecting one of these topic markers over others is not always clear and there is no shortage of cases where native speakers report that alternative topic markers could equally well be used. The richness of the topic marking system, together with difficulty of identifying the distinctive properties of each topic marker, poses considerable challenges for any account of the actual usage of Shanghainese topic markers.

We rely on a corpus of Shanghainese that incorporates a variety of genres in order to collect usage-based data relating to the topic markers. A usage-based (i.e., in effect, a corpus-based) approach is in keeping with the overall shift towards more empirically oriented research within linguistics, but would appear to be especially appropriate, indeed necessary, in the present context where native speaker intuition alone seems unable to provide a fully satisfying explanation of the topic marking system. We do rely on native speaker intuition for some aspects of the analysis of our data, and the first author of this paper (WH) is indeed a resident of Shanghai and a native speaker of Shanghainese. But it is the corpus, reflecting actual usage in a variety of social contexts, which is the empirical foundation of our analysis, rather than constructed, decontextualized examples.

Adopting a corpus-based approach does not imply any specific method of analysis. In the present study we take a decidedly quantitative approach utilizing a number of statistical techniques. A more quantitative approach is particularly apt when the phenomenon under investigation resists any simple analysis based on native speaker's judgments about which topic marker is most appropriate for a given context. In circumstances such as these, it is only through a methodical treatment of the quantitative facts that the influences of competing factors can be properly assessed. We provide therefore a description of the use of topic markers which takes into account a whole set of variables, the core idea behind a multifactorial analysis. Following Arppe (2008), we recognize the usefulness of a general three-tiered framework on how to proceed, consisting, in turn, of univariate, bivariate, and multivariate stages. For the purposes of this paper, however, we focus on a multivariate approach, revealing behaviors of the Shanghainese topic markers that would otherwise be quite elusive. In addition, we make use of predictive methods that go beyond merely observing patterns within the corpus; rather, the methods lead to predictions which assign probabilities to the choice of a topic marker based on a combination of variables.

## 2. The corpus

The language under study in this paper is Shanghainese (上海闲话 *zɑ̃*$^{23}$*hɛ*$^{34}$*ɦɛ*$^{23}$*ɦo*$^{23}$ in Shanghainese pronunciation), a dialect of the Northern Wu branch of the Sinitic language family, spoken in the city of Shanghai and surrounding regions. We defer to the common tradition of referring to Shanghainese as a "dialect", though one needs to be aware that Chinese dialects are more akin to what linguists would call distinct, but genetically related, languages. Shanghainese once served as the regional lingua franca of the entire Yangtze River Delta region and contains elements drawn from the different parts of the Northern Wu area (southern Jiangsu, northern Zhejiang). There are no official statistics available for the number of native speakers of Shanghainese (as distinct from the number of residents of the city or the greater Shanghai region). *The Encyclopedia of Shanghai* (2010: 403) reported 18.8846 million residents in the greater Shanghai area by the end of 2008, a number which includes permanent residents (who might be expected to be speakers) and those who have taken up residence more recently (who are most likely not speakers). This figure would be clearly an over-estimation of the number of proficient speakers of Shanghainese in the Shanghai region, a number we would estimate to be around 13-14 million.

Shanghainese exists almost exclusively as a spoken form of communication, though there exists some literature in the Shanghainese dialect written in Chinese characters from the Ming (1368–1644) and Qing dynasties (1644–1912). The corpus used as the basis for this study of contemporary Shanghainese dialect is, accordingly, made up of transcribed spoken language, based on language data collected since 2008 in downtown Shanghai and in Edmonton, Canada, as well as data recorded from public or broadcast performances. The corpus contains 128,565 word tokens, where "word" refers to a linguistic unit represented by one or more syllables in speech or characters in the Chinese writing system. The data was drawn from four sub-categories or genres: monologue, interview, script, and conversation. A brief description of each of these sub-categories is given in (1). Between them, the sub-categories represent a good variety of usage of the Shanghainese dialect, including both relatively formal styles (monologue and script) and informal styles (interview and conversation). The corpus excludes some categories which, though interesting in their own right, were deemed to be outside the scope of the present study, e.g., the language employed in performances of Shanghainese opera.

(1) a. *Monologue*. 21 files, 19 of which are based on single speakers invited to talk in Shanghainese, recorded in everyday settings. In some cases, speakers were prompted by a small story-like text and invited to more or less retell the story in their own words. Another 2 files are based on single speakers participating in Shanghainese talk shows in 2009. Total word count: 47,663.

b. *Interview*. 5 files, 4 of which are based on interviews conducted by the researcher with some prepared general questions, recorded in Edmonton, Canada. The fifth file is based on an interview-style show called *Three Happy Brothers* broadcast on television in 2010. Total word count: 31,251.

c. *Script*. 23 files, the largest number of individual files for any sub-category. 16 are based on movies and cartoons which have been dubbed into Shanghainese. The others are transcriptions of various scenes from television shows in which Shanghainese was the original language of the production, for example the *Old Uncle* series first broadcast in 2004. Total word count: 20,942.

d. *Conversation*. 2 files, based on conversations between local Shanghainese speakers in downtown Shanghai, recorded in 2009. Total word count: 28,709.

The spoken data was transcribed in Chinese characters appropriate for Shanghainese, as codified in the dictionary of Qian et al. (2007). Transcribing in Chinese characters has the advantage of making the transcription phase relatively straightforward for someone accustomed to typing Chinese characters using familiar input methods. It means, too, that someone able to read Mandarin written in Chinese characters will have some sense of the meaning associated with the same characters when used to represent Shanghainese. However, this kind of exercise – reading the Chinese characters (intended to represent Shanghainese) as though the text represented Mandarin – can be misleading and is not an entirely reliable way of establishing the meaning behind the characters, read as Shanghainese. We also include a broad phonetic transcription, again following Qian et al. (2007), based on pronunciations of morphemes/words spoken in isolation. Tones in this system are represented by the superscripted "tone letters" indicating the location of the beginning, middle, and end of the tone pattern on a tonal scale. Atonal, or "neutral tone", morphemes, particularly relevant to topic markers, have no tone letters in their transcription. While this choice of transcription style leads to representations which are perhaps not very pleasing to the eye and not so easy to read off for most speakers of the language, it has the advantage of deterring readers from treating the example utterances as simply instances of (strange) Mandarin. In any case, our representations allow readers with some knowledge of IPA to read the Shanghainese examples. For the purposes of processing the data in R and presenting the statistical results, a simpler romanization of the five topic markers was used, as explained in Section 4.

## 3. Topic-comment structures

In this study, *topic* is understood with reference to topic-comment structure. In the simplest formulation, a topic is the initial structural element of a sentence (or utterance in the context of spoken language) which specifies what the sentence is about, with *comment* being the remaining structural element which provides comment on the topic. We follow the conventional understanding of topic-comment structure, qualified in the following way, as proposed in Han (2010: 42): *topic structure* and *comment structure* refer to the two basic parts of a sentence; the topic structure, in turn, consists of a topic (the head of the topic structure) which may be introduced (in English) or followed (in Shanghainese) by a *topic marker*.[1] Within the theory of Systemic Functional Grammar, which has found a special popularity among Chinese linguists, the topic structure corresponds to the *theme* and the comment structure corresponds to the *rheme* (Halliday 1985: 39). As an example, the sentence *As for the wedding guests, the bride and bridegroom should be consulted* contains a topic structure (*as for the wedding guests*) and a comment structure (*the bride and bridegroom should be consulted*). The topic structure of the sentence contains a topic marker (*as for*), followed by the topic (*the wedding guests*). "What the sentence is about" is too vague to qualify as an acceptable definition for most linguists, though the (equally vague) term *aboutness* has found its way into the literature as an expedient way of capturing the essential characteristic of a topic (e.g., Reinhart 1981; Gundel 1985). As used here, a topic exists at the utterance/sentence level and must be distinguished from other uses of the term which appeal to aspects of the larger communicative event. This latter approach gives rise to various other understandings of topic, e.g., as a broad, discourse-based concept (Schiffrin 1992), as relevant shared information (cf. Chafe 1976; Copeland and Davis 1983; Lambrecht 1988), as background knowledge for successful communication (Tomlin 1985), among others.

Topic, in the sense being used here, must also be distinguished from a syntactic subject. The sentence cited above, for example, has *wedding guests* as the main substantive element, or head, of the topic; in addition the sentence contains a subject phrase, *the bride and bridegroom*. Topics may share some properties with syntactic subjects, e.g., subjects often introduce what a sentence is 'about' just as topics do; subjects in some languages occur at the beginning of a sentence just as topics do. Other properties of topics, however, make them unlike syntactic subjects: topics do not typically function as an argument to the predicate in the comment and topics do not typically enter into morphosyntactic agreement with the predicate in the comment. Clearly, topic and subject are related in some interesting, even fundamental, ways but are nevertheless distinguishable (hence the appropriateness of the title *Subject and Topic* of Li's 1976 volume). The notions of 'topic-prominence' and 'subject-prominence' (cf. Li and Thompson 1976) have been helpful to linguists in thinking about the ways in which languages may be located along a continuum in terms of the basic structure of a sentence. Members of the Chinese language family, including Shanghainese, can be safely described as topic-prominent languages.

The Shanghainese example in (2) illustrates a number of the typical properties associated with topic-comment structures in Mandarin and Chinese dialects (cf. Li and Thompson 1976): the (underlined) topic structure occurs in the sentence-initial position; a (lexical) topic marker occurs at the right edge of the topic phrase; a pause (indicated here by a comma) is found after the topic structure; the topic itself, *ka*$^{53}$*na*$^{23}$*da*$^{23}$ 'Canada', is definite. Note, too, that the head of the topic does not enter into any obvious argument slot within the predicate. One can translate the sentence by construing the semantic role of the topic head as specifying the location of the universities ('in Canada') as in the free translation provided in (2). One could equally well construe the topic as a kind of modifier of *ɦo*$^{ʔ12}$*da*$^{23}$ 'university' ('Canadian universities'). Or one could construct a kind of topic-comment structure in the English translation ('When it comes to Canada, there's just a few universities, right, compared with the U.S.'). This looseness in the way in which the topic is linked to the comment is typical of topic-comment structures and contrasts with the tighter bond that exists between a syntactic subject and its verb.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| (2) | 加拿大 | 末, | 就 | 搿 | 几所 | 学堂 | 对伐, |
| | *ka*$^{53}$*na*$^{23}$*da*$^{23}$ | *mə*$^{ʔ}$, | *ʑiɤ*$^{23}$ | *gə*$^{ʔ12}$ | *tɕi*$^{34}$*su*$^{34}$ | *ɦo*$^{ʔ12}$*da*$^{23}$, | *tɛ*$^{34}$*va*$^{ʔ12}$, |
| | Canada | TM | just | these | few | university | right |

| | | |
|---|---|---|
| 美国 | 学堂 | 多 |
| *mε$^{23}$go$^{ʔ12}$* | *ɦo$^{ʔ12}$da$^{23}$* | *tu$^{5}$* |
| U.S. | university | many |

‘There are just a few universities in Canada, right, compared to the large number of universities in the US.’ (Inter004)

There are a number of interesting, less typical features of topic-comment structures to be found in Shanghainese. It is possible for some topic markers to *precede* their topics, an ordering of the elements of the topic structure that is quite atypical for Chinese (see Li, L. 2010; Li, Y. 2010; Xu 2010). The relevant topic markers in Shanghainese involve either the *y*$^{35}$ (关于) or *ȿuo*$^{55}$ (说) morphemes: *kuan*$^{55}$ *y*$^{35}$ (关于), *tuei*$^{51}$ *y*$^{35}$ (对于), *tȿʅ*$^{51}$ *y*$^{35}$ (至于), *iau*$^{51}$ *ȿuo*$^{55}$ (要说) and *uo*$^{214}$ *ȿuo*$^{55}$ (我说). While interesting in their own right and deserving greater attention than they have received so far, these less typical topic-marking structures fall outside the scope of the present paper. It is possible, too, for the topic structure in Shanghainese to include more than one sequence of topic + topic marker (cf. Han 2010: 55-76); it can also show an unusual kind of repetition or “copying” of topic structures (cf. Xu and Liu 1998 and Liu 2004). One unusual feature which we do incorporate into our analysis below is that the comment structure can itself appear to contain a topic structure, with a topic marker as part of the comment structure, as illustrated in (3a) and (3b). In (3a), the topic marker *mə*$^{ʔ}$ (末) appears twice (shown in bold), once at the right end of the topic structure and again as part of what we construe as the comment structure. The topic structure in (3a) is itself complex in that it contains a kind of conjunctive structure equivalent to “on the one hand…, on the other hand…”, but the whole conjunctive structure was felt to constitute the topic structure by our native-speaker co-author. (3b) also has a second topic marker within the comment structure, though here different topic markers are used: *mə*$^{ʔ12}$ (嚜) vs. *mə*$^{ʔ}$ (末). In these examples, we take the material after the first topic marker to be like a comment on the material preceding the comma, so we analyze the whole as topic structure + comment structure, albeit with a topic marker (perhaps functioning here more as an emphatic marker) appearing in the comment structure. There could certainly be alternative linguistic analyses of these repeated topic structures as the examples in (3). For the purposes of the present study, it is only necessary that we recognize and label such structures in some distinctive manner (as we do in Section 4).

(3) a.

| | | | | | |
|---|---|---|---|---|---|
| 我 | 一面 | 听， | 是伐， | 一面 | **末**， |
| *ŋu$^{23}$* | *iɪ$^{ʔ55}$mi$^{23}$* | *tʰin$^{53}$,* | *zɿ$^{23}$va$^{ʔ12}$,* | *iɪ$^{ʔ55}$mi$^{23}$* | ***mə$^{ʔ}$**,* |
| I | on one hand | listen | right | on the other | TM |
| 搿 | 辰光 | 老早 | **末** | 碌起来 | 勒 |
| *gə$^{ʔ12}$* | *zən$^{23}$kuã$^{53}$* | *lɔ$^{23}$ʦɔ$^{34}$* | ***mə$^{ʔ}$*** | *lo$^{ʔ12}$tɕʰi$^{34}$lε$^{23}$* | *la$^{ʔ12}$* |
| that | time | quite early | TM | wake up | SFP |

‘I listened; meanwhile, because it was still early, I got up.’ (Mono014)

b.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 勿是 | 老 | 好个 | **嚜**， | 囡儿 | **末** | 也 | 照顾 | 着 |
| *və$^{ʔ12}$zɿ$^{23}$* | *lɔ$^{23}$* | *hɔ$^{34}$gə$^{ʔ12}$* | ***mə$^{ʔ12}$**,* | *nø$^{23}$ŋ$^{23}$* | ***mə$^{ʔ}$*** | *ɦa$^{23}$* | *ʦɔ$^{34}$ku$^{34}$* | *za$^{ʔ12}$* |
| not | very | good | TM | daughter | TM | also | care | SFP |

‘Isn’t it good that you can take care of your daughter?’ (Conv002)

As already mentioned, Shanghainese is particularly rich in topic markers, the most common ones being: *mə*$^{ʔ}$ (末/麼), *ne*$^{ʔ}$ (呢), *zɿ* (是), *tɔ* (倒), *a*$^{ʔ}$ (也), *a* (啊), *tɔ*$^{21}$*zɿ* (倒是), *tɛ*$^{34}$*va*$^{ʔ}$ 伐) and *zɿ*$^{12-22}$*va*$^{ʔ}$ (是伐).[2] Xu and Liu draw attention to "sole-purpose" Shanghainese topic markers, i.e., the ones used uniquely in the function of marking topics (cf. Han 2010: 78-87) in contrast to those topic markers used for functions other than topic marking (Xu and Liu 2007: 78-80). The existence of these sole-purpose topic markers, they argue, makes Shanghainese a more topic-prominent variety of Chinese than Mandarin, quite apart from the larger choice in topic markers available in Shanghainese.

We opted to concentrate on the five most frequently used topic markers in our corpus, as determined by total occurrence of tokens. These turned out to be *ne*$^{ʔ}$ (呢) with a frequency of 1,117, *a* (啊) 749, *mə*$^{ʔ}$ (末) 687, *zɿ* (是) 305, and *mə*$^{ʔ12}$ (嚜) 152. Some brief background comments on each of these topic markers are in order (cf. Chu [1987: 218] for the early history of these forms). *mə*$^{ʔ}$ (末) and *mə*$^{ʔ12}$ (嚜) are specifically Shanghainese topic markers, with the latter not previously noted in the literature (cf. Note 2). There is, importantly, a tonal difference distinguishing the two topic markers and, as we will see later, they are associated with different preferences for some of the variables we examine. The remaining three topic markers have some currency outside of Shanghainese and have more presence in the history of Chinese. The topic marker *a* (啊) would appear to be the topic marker with the most general currency, being used in many Chinese dialects and Mandarin (and hence, feels most "formal" when used in Shanghainese, according to our Shanghainese-speaking co-author, WH). *ne*$^{ʔ}$ is already attested in Archaic Chinese (Chu 1987: 218) as a topic marker, while *zɿ* is the youngest of the five topic markers, having evolved during the Ming Dynasty from earlier uses as a copula and as a focus marker. In addition to being specific to Shanghainese, *mə*$^{ʔ}$ and *mə*$^{ʔ12}$ are also used *only* as topic markers in Shanghainese (as mentioned in the preceding paragraph). The different historical profiles of the topic markers have a bearing on the attraction or repulsion of the markers to the four genres in the corpus, a point we return to later in our analysis.

As a way of deconstructing the elusive idea of *aboutness* of topics as it applies to Shanghainese, we may distinguish five kinds of constructional meaning associated with Shanghainese topic-comment structures. Each of the topic markers studied in this paper can occur in topic-comment structures associated with any one of these five meanings and it was on the basis of examining the corpus data that we were led to the classification in (i)-(iv). As we define them, the constructional meanings highlight the most striking semantic or pragmatic aspect of the topic-comment structure, located in either the topic structure or comment structure. It should not be thought that it is the topic marker per se that conveys all of the semantic/pragmatic effect; the meanings in (i)–(v) are associated with the whole topic-comment construction.

(i) *Introductory* ('*given-new*') *meaning*

The *introductory* meaning is found in the most neutral, or unmarked, kind of information structure carried by topic-comment structures. It refers to the introduction of new and highlighted information in the comment structure, relating to a known or given topic. In (4), *mə*$^{ʔ}$ serves to mark a known, unemphatic topic *ɦi*$^{23}$ 'he'. The comment introduces new information about the topic, as is typical of all topic-comment structures. Crucially, however, it is the information in the comment which is the communicatively more salient part of the sentence (as ascertained by consideration of the larger context in which the utterance occurs).

| | | | | | |
|---|---|---|---|---|---|
| (4) | 伊 | 末, | 已经 | 老油条 | 唻 |
| | *ɦi*$^{23}$ | *mə*$^{ʔ}$*,* | *i*$^{23}$*tɕin*$^{53}$ | *lɔ*$^{23}$*ɦiɤ*$^{23}$*diɔ*$^{23}$ | *lɛ*$^{23}$ |
| | he | TM | already | sophisticated like a wily old bird | SFP |

‘He’s already sophisticated like a wily old bird!’ (Script006)

(ii) *Emphatic meaning*

An *emphatic* meaning is associated with topic-comment structures in which the topic, whether it is new or given information, is emphasized more than the comment. Again, the larger context of the utterance is important in determining which parts of the utterance are regarded as emphasized. In (5), *mə*$^{ʔ}$ marks the emphasized topic *gə*$^{ʔ12}$*tsoŋ*$^{34}$ *ku*$^{53}$ ‘this type of song’, and the italics in the free translation are meant to give some sense of the emphatic quality of the topic.

| | | | | | | |
|---|---|---|---|---|---|---|
| (5) | 搿种 | 歌 | 末, | 最 | 直接 | 唻 |
| | *gə*$^{ʔ12}$*tsoŋ*$^{34}$ | *ku*$^{53}$ | *mə*$^{ʔ}$*,* | *tsø*$^{34}$ | *zə*$^{ʔ12}$*tɕiɪ*$^{ʔ55}$ | *lɛ*$^{23}$ |
| | this type | song | TM | most | direct | SFP |

‘It is *this* type of song that is the most direct.’ (Script011)

(iii) *Contrastive meaning*

The *contrastive* meaning refers to topic-comment structures in which the topic stands in specific contrast to other known information from either preceding linguistic context (“co-text”) or from the situational context. In (6), the preceding linguistic context concerns the amount of food to be eaten in the morning. The temporal adverbial *ɦiɑ*$^{23}$*dɤ*$^{23}$ ‘evening in the (underlined) topic structure is strongly contrastive with the reference to the earlier time *tsɔ*$^{34}$*zən*$^{23}$*dɤ*$^{23}$ ‘morning’, hence the topic-comment structure is categorized as having contrastive meaning.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| (6) | 我伲 | 人 | 呢， | 早晨头 | 量 | 要 | 好， |
| | *ŋu*$^{23}$*ɲi*$^{23}$ | *ɲin*$^{23}$ | *ne*$^{ʔ}$*,* | *tsɔ*$^{34}$*zən*$^{23}$*dɤ*$^{23}$ | *liɑ*$^{\sim 23}$ | *iɔ*$^{34}$ | *hɔ*$^{34}$*,* |
| | we | people | TM | morning | quantity | should | be.good |
| | 夜头 | 呢, | 要 | 吃 | 得 | 少 | |
| | *ɦiɑ*$^{23}$*dɤ*$^{23}$ | *ne*$^{ʔ}$*,* | *iɔ*$^{34}$ | *tɕyə*$^{ʔ55}$ | *tə*$^{ʔ55}$ | *sɔ*$^{34}$ | |
| | evening | TM | should | eat | PARTICLE | little | |

‘We should eat a lot in the morning, but in the *evening* we should eat less.’
(Mono014)

(iv) *Conditional meaning*

The *conditional* meaning attaches to topic-comment structures in which the topic functions as the condition or prerequisite for the event or state of affairs described by the comment. In (7), for example, the topic presents the condition (being successful in running a business) for further promotion.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| (7) | 侬 | 做辣 | 好 | 末, | 下趟 | 拨 | 侬 | 两只柜台， |
| | *noŋ*$^{23}$ | *tsu*$^{34}$*la*$^{ʔ12}$ | *hɔ*$^{34}$ | *mə*$^{ʔ}$*,* | *ɦo*$^{23}$*tʰã*$^{34}$ | *pə*$^{ʔ55}$ | *noŋ*$^{23}$ | *liã*$^{23}$*tsə*$^{ʔ55}$*tɕy*$^{34}$*dɛ*$^{23}$*,* |
| | you | do | well | TM | next time | give | you | two stalls, |

| | | |
|---|---|---|
| 拨 | 侬 | 一爿店 |
| *pə*$^{ʔ55}$ | *noŋ*$^{23}$ | *iɪ*$^{ʔ55}$*pɛ*$^{34}$*ti*$^{34}$ |
| give | you | a branch |

‘If you do well this time, you will be appointed in charge of two stalls, even a branch.’ (Conv002)

(v) *Counter-expected meaning*

*Counter-expected* meaning is found in cases where the comment presents information which is contrary to expectation, negating any presuppositions or conversational implicatures previously established, as in (8).

(8)

| | | | | | |
|---|---|---|---|---|---|
| 但是 | 伊 | 呢, | 倒 | 勿是 | 一个 |
| *dɛ*$^{23}$*zɿ*$^{23}$ | *ɦi*$^{23}$ | *ne*$^{ʔ}$, | *tɔ*$^{34}$ | *və*$^{ʔ12}$*zɿ*$^{23}$ | *iɪ*$^{ʔ55}$*gə*$^{ʔ12}$ |
| however | he | TM | against expectation | not | a |

| | |
|---|---|
| 守财奴, | 一毛勿拔个 |
| *sɤ*$^{34}$*zə*$^{53}$*nu*$^{23}$, | *iɪ*$^{ʔ55}$*mɔ*$^{23}$*və*$^{ʔ12}$*ba*$^{ʔ12}$*ə*$^{ʔ12}$ |
| miser, | stingy |

‘However, he’s actually not a stingy miser.’ (Mono019)

Clearly, some subjective decision-making lies behind the determination of which semantic or pragmatic meaning is most salient in these structures, since the corpus is not annotated for semantic/pragmatic features. What we have called the introductory meaning is arguably present, to varying degrees, in all five of these categories. So, for example, in (5), illustrating the emphatic meaning, there is an underlying ‘given-new’ pragmatic structure, but the emphatic component of the meaning is taken to be the most salient part of the conveyed meaning. In a similar way, a counter-expected meaning, as in (8), includes a kind of contrast between what is asserted and what the expectation is, but we categorize the example in the more specific way as counter-expected.

## 4. Methodology

Our approach involves treating the five most frequent topic markers identified in Section 3 as the dependent variable, with various other features of the context of usage of the topic markers as the independent variables. In what follows, we explore how the behaviors of the topic markers can be understood and explained in terms of these other variables. We identified five independent variables that appeared to us to potentially have some bearing on the choice of topic marker. These variables, or “factors”, are summarized in (9). For each variable, there is a number of sub-categories, or unordered “levels” or “categories”. Undoubtedly, it would have been revealing to have included demographic variables such as age, gender, and education level achieved. However, the relevant information for these categories had not been included systematically as part of the metadata for the entire corpus and so the analysis did not include such factors.

(9) a. Numeric variable: *length of topic*, as measured by the number of syllables in the topic constituent, excluding the topic marker itself[3]: 1 – 10, where 10 stands for 10 or more syllables

b. Factor: *syntactic category of the topic*
5 Levels: NOMINAL (NOM), VERBAL (VERB), ADJECTIVAL (ADJ), ADVERBIAL (ADV),

CLAUSAL (CLAUSE)

c. Factor: *main function of the topic-comment structure*
5 Levels: INTRODUCTORY (INTR), EMPHATIC (EMPH), CONTRASTIVE (CONT), CONDITIONAL (COND), COUNTEREXPECTED (COUNTER)

d. Factor: *comment type*
5 Levels: CLAUSE (CLAUSE), PHRASE (PHRASE), TAG (FINALTAG)[4], COMMENT STRUCTURE CONTAINING SAME TOPIC MARKER as used in initial topic, as in 3a above (SAMEMRKR), COMMENT STRUCTURE CONTAINING DIFFERENT TOPIC MARKER as used in initial topic, as in 3b above (DIFFMRKR)

e. Factor: *genre*
4 Levels: MONOLOGUE (MONO), INTERVIEW (INTER), SCRIPT (SCRIPT), CONVERSATION (CONV)

We retrieved 100 random lines from the corpus for each topic marker in the initial topic structure for a total of 500 lines, where each line represents a whole utterance. A spreadsheet was used to list these lines, with each line coded for each of the five factors. The result of all this is a "dataframe" as recognized by the statistical programming language R and the basis for the statistical calculations below. For processing the data in R it was more convenient to represent the five topic markers in a broad romanized transcription and we will henceforth use these simplified transcriptions, without the accompanying character: *ne*$^{ʔ}$ (呢 ) > *ne*, *a* (啊)  > *a*, *mə*$^{ʔ}$ (末) > *ma*, *zɿ* (是) > *zi*,  and *mə*$^{ʔ12}$ (嚜) > *mo*.

## 5. Statistical analysis

Our main intention in this study is to carry out a multivariate analysis of the Shanghainese data, using several functions in the `polytomous` package (Arppe 2013) in *R*, the public-domain statistical programming environment (R Core Development Team, 2012). Indeed, the present study serves as a way of introducing readers to this package.

Before we embark on the multivariate analysis, we would like to briefly draw attention to the possibilities for various kinds of univariate (and bivariate analyses) within the `polytomous` package. One might, for example, be interested in exploring the over-representation or under-representation of, say, the INTRODUCTORY FUNCTION with each topic marker. As a simple way of inspecting this single level of the FUNCTION factor, one might want to consider a cross-tabulation like that in Table 1. In this table, the occurrences of each topic marker *with* this function (shown in the first row) are contrasted with the occurrences of each topic marker *without* this function, i.e. any of the other FUNCTION categories (in the second row). The `chisq.posthoc()` function in the `polytomous` package offers many options to the researcher wishing to systematically explore distributions at this level.  One can, for example, use this function to display the statistically significant instances of over- and under-representation in the top row of Table 1 as in Table 2. In this kind of output, pluses, minuses and zeros (+/-/0) are used to show the significant divergences, or lack thereof, based on the standardized Pearson residuals (cf. Agresti 2002: 78-80; Arppe 2008: 75-84). We can now notice that the INTRODUCTORY FUNCTION occurs significantly more than expected with *zi* and significantly less than expected with *ne*, but for the three other topic markers *a*, *ma* and *mo* the individual divergences do not surpass the prescribed threshold values either way. We invite the reader to explore the full range of univariate and bivariate functions available in the `polytomous` package, as illustrated in the R vignette (Arppe 2013; Arppe, Han, and Newman in prep.).

Table 1. Cross-tabulation of the INTRODUCTORY vs. other FUNCTIONs across the five topic markers (raw frequencies)

| | zi | a | ma | mo | ne |
|---|---|---|---|---|---|
| INTRODUCTORY FUNCTION | 67 | 43 | 36 | 33 | 18 |
| ¬ INTRODUCTORY FUNCTION | 33 | 57 | 64 | 67 | 82 |

Table 2. Preferences for the distribution of the INTRODUCTORY FUNCTION among the five topic markers, corresponding to Table 1, as determined by the `chisq.posthoc()` function

| | zi | a | ma | mo | ne |
|---|---|---|---|---|---|
| INTRODUCTORY FUNCTION | + | 0 | 0 | 0 | - |

Among various multivariate statistical methods for more than two possible outcomes, as is the case with the five topic markers here, *polytomous logistic regression* analysis (see, for example, Hosmer and Lemeshow 2000: 260-287; Arppe 2008:113-116) appeared to be the most attractive approach. As a *direct probability model* (Harrell 2001: 217), polytomous, as well as binary, logistic regression yields probability estimates, corresponding to the expected proportions of occurrences, conditional on the values of the explanatory variables that have been selected for inclusion in the model. This characteristic fits well together with prior linguistic research (e.g., Featherston 2005; Bresnan et al. 2007; Arppe and Järvikivi 2007), from which we know that in practice individual features or sets of features are *not* observed in corpora to be categorically matched with the occurrence (in a corpus) of only one word/construction in some particular synonymous and no others. While one topic marker among the possible variants may be by far the most frequent for some particular context, others do also occur, albeit with often a considerably lower relative frequency. Furthermore, with respect to the weighting of individual variables in polytomous logistic regression, the parameters associated with each variable have a natural interpretation in that they reflect the increased (or decreased) *odds* of a particular outcome occurring, when the particular feature is present in the context, with all the other explanatory variables being equal. The exact meaning of the odds varies depending on which practical heuristic has been selected, and can involve, for example, a contrast of an outcome category with all the rest or with some baseline category.

There are a number of heuristics for implementing polytomous logistic regression, which are all based on the splitting of the polytomous setting into a set of dichotomous cases, to each of which a corresponding binary logistic regression model can then be applied and fitted either simultaneously or separately. These heuristics are presented and their characteristics discussed from the linguistic perspective in Arppe (2008: 113-116, 119-125; see also Frank and Kramer 2004). In order to get both topic-marker-specific parameters for the selected explanatory features, without having to select one topic marker as a baseline category, and probability estimates for the occurrences of each topic marker, we found the *one-vs-rest* heuristic (Rifkin and Klautau 2004; Arppe 2008: 120-121; 2009) to be the most appealing. This methodological choice is facilitated by the observation that its performance does not significantly differ from that of the other heuristics (Arppe 2008: 198-201). The one-vs-rest model concerning the topic markers was fitted using the `polytomous` function in the `polytomous` package (Arppe 2013). The predictors used in the model are the features introduced in Section 4. One must note that for each categorical value that has fully complementary values covering the entire dataset, one such class/category needs to be designated as the default value in order to avoid exact collinearity, being typically the most frequent or prototypical one or that feels least surprising to the analyst. In the present case, these were the INTRODUCTORY FUNCTION, NOMINAL TOPIC-PART-OF-SPEECH, CLAUSAL COMMENT TYPE, and MONOLOGUE GENRE. (10) is the complete summary output from applying the `polytomous` function to our dataset, based on these default values.

(10) Summary of results from the `polytomous` function in *R*. Estimated odds for explanatory features in favor of or against the occurrence of the topic marker outcomes are shown under Odds; non-significant odds (P<0.05) are shown in parentheses.

```
> print(summary(polytomous(TOPIC_MARKER ~ TOPIC_LENGTH + TOPIC_POS + FUNCTION
+ COMMENT_TYPE + GENRE, shanghainese)), max.print=NA)

Formula:
TOPIC_MARKER ~ TOPIC_LENGTH + TOPIC_POS + FUNCTION + COMMENT_TYPE +
    GENRE

Heuristic:
one.vs.rest

Odds:
                       ne          a           mo          zi          ma
(Intercept)            0.08782     0.2628      0.009559    5.804       0.2585
COMMENT_TYPEPHRASE     (0.568)     (0.7328)    (1.42)      (1.283)     (1.42)
COMMENT_TYPETFINALTAG  (1.261)     (0.5884)    (1.22)      (0.855)     (0.9522)
COMMENT_TYPESAMEMRKR   (0.3588)    (2.456)     (1.632)     (0.5519)    (0.5734)
COMMENT_TYPETDIFFMRKR  (0.8791)    (1.634)     (0.6193)    (0.4156)    (1.1)
FUNCTIONCOUNTER        9.425       (0.8001)    (0.3065)    0.1788      (1.151)
FUNCTIONCOND           7.556       0.3025      (1.313)     0.07156     (1.055)
FUNCTIONCONT           12.87       0.281       0.09028     0.02395     2.697
FUNCTIONEMPH           2.833       (1.364)     (0.6123)    0.525       (1.07)
GENRECONV              0.3315      0.03177     4.286       (0.8568)    4.046
GENREINTER             (0.8097)    0.1964      9.254       (1.043)     (0.5909)
GENRESCRIPT            (0.2833)    (1.688)     (3.219)     (0.4916)    (1.011)
TOPIC_LENGTH           1.185       (1.122)     1.437       0.5414      0.8104
TOPIC_POSADJ           (0.2623)    (0.5769)    12.54       0.1827      (1.462)
TOPIC_POSADV           (1.629)     (1.119)     (0.2926)    (0.5171)    (1.307)
TOPIC_POSCLAUSE        0.2126      (0.9687)    2.99        (0.7534)    (2.192)
TOPIC_POSVERB          0.4591      (1.815)     2.709       0.2649      (1.875)

Null deviance:             1609  on  2500  degrees of freedom
Residual (model) deviance:  1191  on  2415  degrees of freedom

R2.likelihood:  0.26
AIC:            1361
BIC:            1719
```

We start by looking at the overall performance and fit of the model in terms of two measures. The first statistic, $R_L^2$ (the `R2.likelihood` value in (10)), is an indicator of how well a logistic regression model fits with the actual occurrences in the original data (Hosmer and Lemeshow 2000: 165-166; Arppe 2008: 126-129). This is calculated as a comparison of the probabilities predicted by the model for each actually occurring outcome and the associated feature cluster, against the baseline probability for each outcome class, the latter being simply the topic markers' overall proportions in the entire data. In comparison to the $R^2$ measure used in ordinary linear regression, $R_L^2$ does *not* tell us the proportion of variation in the data that a logistic regression model succeeds in explaining, but $R_L^2$ does allow us to compare the overall fit of different models with varying sets of explanatory variables on the same data. The $R_L^2$ =0.26 for the current model can be considered relatively good for polytomous logistic regression models.[5]

The second measure, *Accuracy*, concerns efficiency in prediction (Menard 1995: 28-30; Arppe 2008: 129-132) and tells us how often overall a prediction is correct, based, in the case of the one-vs-rest heuristic, on a prediction rule of selecting for each context the topic marker receiving the highest probability estimate. The Accuracy value of 0.500 for the current model is in fact an aggregate of the topic-marker-wise Accuracy values, which are quite divergent, favoring *mo* with an Accuracy of 61%, in comparison to the respective values of 57% for *a*, 57% for *zi*, 39% for *ne*, and 36% for *ma*. Underlying the Accuracy values is a cross-tabulation of the originally occurring topic-markers and the predicted ones, presented in Table 3. Rows in Table 3 sum to 100, since there were originally 100 examples of each topic marker to be observed in the database. The main point to note is that the most frequently predicted topic-marker in each column always corresponds to the originally occurring topic marker (as shown by the bold numbers in Table 3), confirming that, overall, the model "gets it right". Moreover, one can scrutinize Table 3 in terms of which topic markers are mistaken for each other, and to what extent. Consider the wrong predictions the model makes for *zi* and *a*. The most frequent wrong prediction for (correct) *zi* is *a*, and conversely the most frequent wrong prediction for (correct) *a* is *zi*, with identical error rates (23% in both cases). Consider, too, the predictions in the *ma* cases in the dataframe. As indicated by the standard deviation in the set of numbers for each row in Table 3, it is *ma* which shows both the least Accuracy (36%) and the least deviation in prediction rates. In other words, *ma* is not strongly predicted by this model and the competition between *zi*, *a*, *mo*, and *ne* is relatively equal in the cases where *ma* was in fact used. One can interpret these facts as indicating (correctly, we believe) a relatively general or 'default' topic marker, a topic marker that is used commonly, but without any particularly strong factor motivating its use. In Table 3 it is *mo* that is most accurately predicted of all the topic markers and it is the marker with the highest standard deviation associated with it. These facts can be seen as further confirmation of the distinctiveness of *mo* vis-à-vis the other markers. Finally, in assessing the Accuracy of a model, one must remember that logistic regression analysis models primarily relative proportions of occurrences rather than categorical selections. Thus, selecting always the topic marker with the highest probability estimate, given a context, masks the fact that the model also assigns some probability to the other topic markers, too, entailing that the model predicts these less likely topic markers as also occurring in that particular context, though with smaller overall proportions.

Table 3. Cross-tabulation of originally occurring topic markers and those predicted by the polytomous logistic regression model. Correct predictions are shown in bold.

| Predicted<br>Observed | zi | a | ma | mo | ne | standard deviation in the predicted values |
|---|---|---|---|---|---|---|
| zi | **54** | 23 | 8 | 6 | 6 | 21.9 |
| a | 23 | **57** | 4 | 8 | 8 | 21.9 |
| ma | 14 | 22 | **36** | 11 | 17 | 9.8 |
| mo | 11 | 16 | 6 | **61** | 6 | 23.3 |
| ne | 12 | 20 | 8 | 21 | **39** | 11.9 |

We now look at the impact of the various explanatory features in the use and choice of the five topic markers, as summarized in the Odds section of (10). We can again look at the results from either the topic-marker-wise or feature-wise perspective, opting now primarily for the former. Generally, we may note that the odds for all categories of COMMENT TYPE are not significant. For the individual topic markers, the aggregate of the default values of the categorical variables significantly increases only the chances of *zi* to occur, reflected in the *Intercept* odds of 5.8:1, whereas for the four other topic markers the odds for the Intercept are significantly against their occurrence. (11) summarizes the key results from the table of Odds in (10).

(11) a. *zi:* While being significantly preferred by the aggregate of default variable values, the chances of *zi* occurring are significantly decreased by the CONTRASTIVE (0.02:1), CONDITIONAL (0.07:1), COUNTEREXPECTED (0.18:1) and EMPHATIC (0.53:1) FUNCTIONs, the ADJECTIVAL (0.19:1) and VERBAL TOPIC-PART-OF-SPEECH (0.26:1), and TOPIC LENGTH (0.54:1).

b. *a*: *a* has no features significantly in its favor, but instead the CONVERSATIONAL (0.03:1) and INTERVIEW (0.20:1) GENREs exhibit significant and strong odds against its occurrence, followed by the CONTRASTIVE (0.28:1) and CONDITIONAL FUNCTIONs (0.30:1).

c. *ma*: The chances of *ma* occurring are significantly increased by the CONVERSATIONAL GENRE (4.0:1), followed by the CONTRASTIVE FUNCTION (2.7:1), whereas TOPIC LENGTH (0.81:1) is slightly but significantly against this particular topic marker.

d. *mo*: In the case of *mo*, the ADJECTIVAL TOPIC-PART-OF-SPEECH is most significantly in favor of its occurrence (12.5:1), followed by the INTERVIEW (9.3:1) and CONVERSATION (4.3:1) GENREs, the CLAUSAL (3.0:1) and VERBAL (2.7:1) TOPIC-PART-OF-SPEECH, and TOPIC LENGTH to a lesser but nonetheless significant effect (1.4:1). However, the CONTRASTIVE FUNCTION (0.09) shows strong significant odds against the occurrence of *mo*.

e. *ne*: the chances of *ne* occurring are mostly increased by the CONTRASTIVE FUNCTION (12.9:1), followed by the COUNTEREXPECTED (9.4:1) and CONDITIONAL FUNCTIONs (7.6:1), with the odds turning more moderate but still significant with the EMPHATIC FUNCTION (2.8:1) and TOPIC-LENGTH (1.2:1). The CLAUSAL TOPIC-PART-OF-SPEECH shows the strongest significant odds (0.21:1) against the occurrence of this topic marker, followed by the CONVERSATIONAL GENRE (0.33:1).

## 6. Probability estimates

In addition to assigning odds for the explanatory variables, as discussed above, another attractive characteristic of a (polytomous) logistic regression model is its ability to provide probability estimates for an outcome, given any possible mix of explanatory variables, representing a set of features present in some context. Like the estimated odds, the accuracy of such probability estimates is naturally dependent on how well the explanatory variables incorporated in the model are able to describe and fit the data they are trained with, as well as to predict instances in new, unseen data, that is, how generally applicable the selected model is. Nevertheless, the probability estimates allow us to effectively rank with a single value the joint effect of a large number of features and their complex interrelationships, which is typically the case with real, natural usage of language. For any combination of the five factors in the context, we are able to determine the probability $P$ of a topic marker given a particular context, i.e., $P(Topic\text{-}Marker|Context)$.[6] In fact, this application is possible with any statistical technique that is probabilistic (or can be interpreted as such), as has been demonstrated earlier on (e.g., Gries 2003) using Linear Discriminant Analysis (LDA) for ranking sentences, representing two constructional alternatives denoting the same meaning, in terms of their *prototypicality* with respect to the two alternatives. Gries' approach effectively merges the concepts of prototype and exemplar by seeing these as manifested primarily in the original sentences (and their constituent properties) in the dataset, and undertakes the ordering of sentences in the data in terms of their prototypicality with respect to the two alternative constructions along a single axis, with the two alternatives at the opposite ends. An alternative approach is to distinguish between the selection of exemplars and the representation of the prototypes (e.g. Divjak and Arppe 2013).

The maximum probabilities assigned for each combination of contextual factors determine the topic marker predicted for that combination and so obviously it is of interest to identify the topic marker most predicted for each context. In addition, though, it is instructive to examine the entire spectrum of probabilities estimated for each topic marker ($T$) in a particular context ($C$), especially since logistic regression analysis models relative proportions of occurrences (in the long run) rather than categorical selections. Figure 1 provides an overview of the probabilities of all topic markers through five density plots, with the probabilities broken down into five bands representing the distributions of the maximum down to the minimum values, as ranked over each sentence. The density plots are helpful in so far as they reveal the overall ranges in each band of probabilities. For a start, one can see that the maximum probability assigned for any topic marker in any context never reaches even close to the theoretical maximum $P(T|C)=1.0$, being rather $P_{max}(T|C)=0.822$, and the predictions are closest to categorical in only 3 (0.6%) instances for which $P_{max}(T|C)>0.8$. The mean probabilities for the top three bands (shown in the upper row of Figure 1) are $P_{max}(T|C)=0.490$ for the maximum band, $P_{max-1}(T|C)=0.258$ for the second highest band, and $P_{max-2}(T|C)=0.143$ for the third highest band. Even the mean of context-wise minimum estimated probabilities is clearly above nil, being $P_{min}(T|C)=0.029$. Quite a few of the contexts can realistically have two or even more outcomes, though preferential differences among the topic-markers remain to varying extents (cf. Hanks 1996: 79; Arppe 2009: 13-14). To give an example of how close some of the predictions for alternative topic markers can be, note that there are 154 (30.8%) cases where the top two probability estimates for the topic marker are $P(T|C)\geq0.3$. In these cases, it would appear that we are dealing with highly interchangeable topic markers.

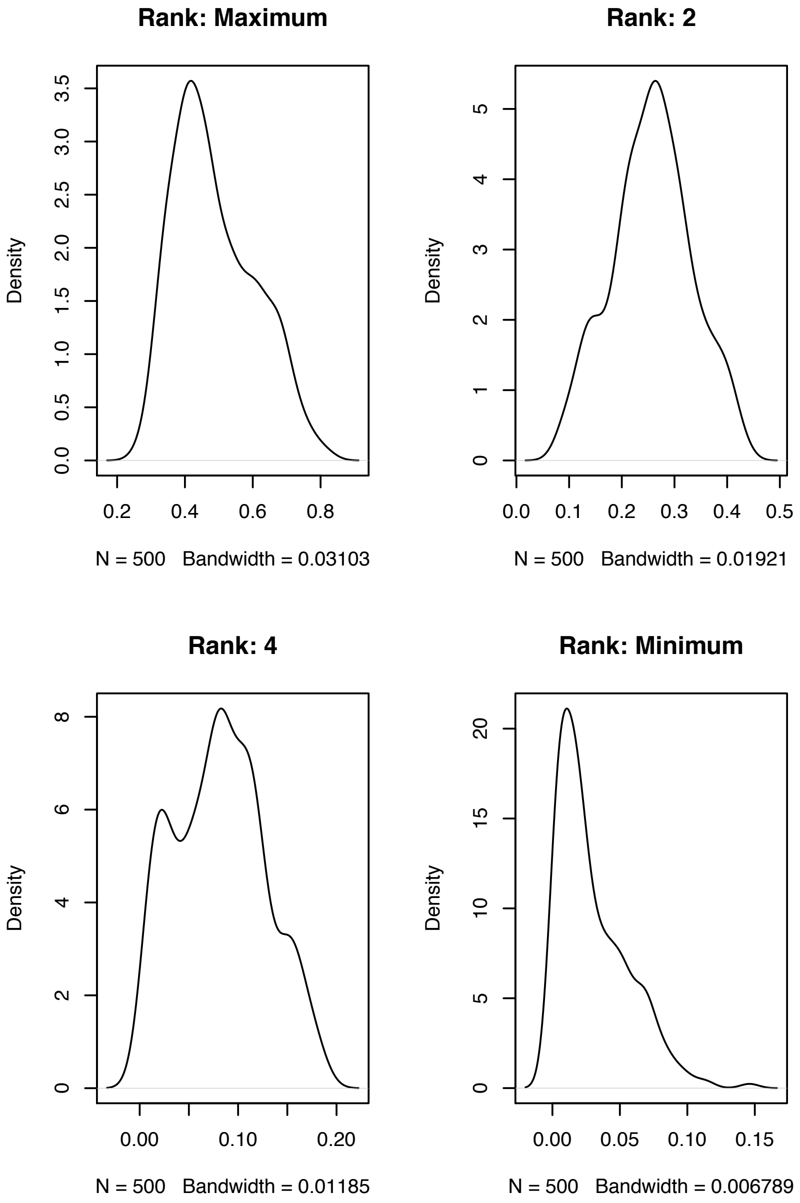

Figure 1. Densities of the distributions of the estimated probabilities by rank order for all instances in the data (*n=500*)

Zooming in on individual sentences in the research corpus, we can observe various scenarios of how the entire estimated probability space (with $\sum P[T|C]=1.0$) can be distributed among the topic markers on the basis of the selected features manifested in each context (cf. Arppe 2008: 237-247). As noted above, there are no cases where the probability distribution approaches a categorical, exception-less choice, so that only one of the topic markers is assigned even close to the maximum possible probability $P(T|C)\approx 1.0$, while the rest receive none (in contrast, for example, to the noticeable though small number of such contexts as observed in Arppe (2008: 239, Table 5.31). Instead, we find contexts with some inherent degree of variation so that while one topic marker is clearly preferred in such circumstances, receiving the highest probability, one or more of the others may also have a real though more occasional chance of occurring to varying degrees. We will illustrate the variation in the profiles of the probability estimates for any one context through four examples below from (12) to (15).

(12) illustrates the case where the primary preference appears quite strong, accompanied by a clear second-best choice. (12) combines the contextual factors of CLAUSAL TOPIC-PART-OF-SPEECH, a CONTRASTIVE FUNCTION, a CLAUSAL COMMENT-TYPE, use in the CONVERSATION GENRE, and a TOPIC-LENGTH of 3. Note that there are, in fact, two topic-comment structures evident in (12b); it is the second instance that is the basis for the estimated probabilities. As can be seen in the breakdown of probabilities for this context, *ma* has the highest such probability at *P=0.822*, with a clear second-best outcome *ne* at *P=0.125*. There are smaller, but non-zero, estimates for *mo* and *zi*, while *a* can be considered practically improbable with *P=0.004*. Our Shanghainese-speaking co-author's intuition was that both *ma* and *ne* would be the two best choices in (12b), noting also that the topic in question (underlined in 12b) forms a parallel structure and a strong contrast with the topic at the beginning of (12b). We had already established that the CONTRASTIVE function is strongly associated with *ma* and the high estimated probability for *ma* in (12a) compared with *ne* is presumably heavily influenced by the CONTRASTIVE function (and the CONVERSATION genre).

(12) a. Probability estimates and context for case #222

| | |
|---|---|
| P(*ne*\|$C_{\#222}$)=0.125<br>P(*a*\|$C_{\#222}$)=0.004<br>P(*mo*\|$C_{\#222}$)=0.034<br>P(*zi*\|$C_{\#222}$)=0.015<br>P(*ma*\|$C_{\#222}$)=**0.822 (predicted correctly)** | Context: topic length 3, clause topic, contrastive function, clause comment-type, conversation genre |

b.

| | | | | | | |
|---|---|---|---|---|---|---|
| 侬 | 高兴 | 末 | 侬 | 就 | 做做， | |
| *noŋ*$^{23}$ | *kɔ*$^{53}$*ɕin*$^{34}$ | *mə*$^{ʔ}$ | *noŋ*$^{23}$ | *ʑiɤ*$^{23}$ | *ʦu*$^{34}$*ʦu*， | |
| you | happy | TM | you | then | do | |
| 勿 | 高兴 | 末， | 就 | 覅 | 做 | |
| *və*$^{ʔ12}$ | *kɔ*$^{53}$*ɕin*$^{34}$ | *mə*$^{ʔ}$, | *ʑiɤ*$^{23}$ | *viɔ*$^{23}$ | *ʦu*$^{34}$ | |
| not | happy | | TM | then | not | do |

'If you're happy then do it; if you're unhappy, then ignore it.' (Conv002)

Another kind of profile is seen in (13), where the context consists of a VERBAL TOPIC-PART-OF-SPEECH, an INTRODUCTORY FUNCTION, a TOPIC-STRUCTURE-WITH-SAME-TOPIC-MARKER COMMENT-TYPE, use in the MONOLOGUE GENRE, and a TOPIC-LENGTH of 4. The primary preference is for *a* at *P=0.688*, weaker than the best choices in (12) but still a clear winner. The other alternatives show smaller, but not insubstantial, probabilities all in the range from *P=0.109* for *ma* down to *P=0.035* for *zi*. Our Shanghainese-speaking co-author felt that, intuitively, *a* was the most viable choice, related in particular to the stacking up of four topic phrases.

(13) a. Probability estimates and context for case #185

| | |
|---|---|
| $P(ne \mid C_{\#185})=0.072$<br>$P(a \mid C_{\#185})=$**0.668 (predicted correctly)**<br>$P(mo \mid C_{\#185})=0.095$<br>$P(zi \mid C_{\#185})=0.035$<br>$P(ma \mid C_{\#185})=0.109$ | Context: topic length 4, verb topic, introductory function, topic structure with same topic marker comment-type, monologue genre |

b.

| | | | | | |
|---|---|---|---|---|---|
| [最高 | 指示, | 搿 | 只要 | 一发布,] | |
| *tsø*$^{34}$*kɔ*$^{53}$ | *tsɿ*$^{34}$*zɿ*$^{23}$, | *gə*$^{ʔ12}$ | *tsə*$^{ʔ55}$*iɔ*$^{34}$ | *iɪ*$^{ʔ55}$*fa*$^{ʔ55}$*pu*$^{34}$, | |
| highest | order | it | only when | issue | |
| 敲锣打鼓 | 啊， | 游行 | 啊， | 造反队 | 啊， |
| *kʰɔ*$^{34}$*lu*$^{23}$*tã*$^{34}$*ku*$^{34}$ | *a*， | *ɦiɤ*$^{23}$*ɦin*$^{23}$ | *a*， | *zɔ*$^{23}$*fɛ*$^{34}$*dɛ*$^{23}$ | *a*， |
| drum beat | TM | parade | TM | rebel force | TM |
| 红卫兵 | 啊， | 对伐， | | | |
| *ɦoŋ*$^{23}$*ɦuɛ*$^{23}$*pin*$^{53}$ | *a*， | *tɛ*$^{34}$*va*$^{ʔ12}$， | | | |
| red guard | TM | right | | | |
| 侪 | 出来 | 迭个 | 是 | 跳舞 | |
| *zə*$^{53}$ | *tsʰə*$^{ʔ55}$*lɛ*$^{23}$ | *diɪ*$^{ʔ12}$*gə*$^{ʔ12}$ | *zɿ*$^{23}$ | *tʰiɔ*$^{34}$*vu*$^{23}$ | |
| all | come out | it | be | dance | |

'[Only when the highest orders are issued,] there'll be the drum beating, a parade, the Rebel Forces, Red Guards, right, they will all come out to dance.' (Mono 005)

(14) illustrates an example where the estimated highest probability does not coincide, in fact, with the topic marker selected by the speaker. It is a feature of the modeling approach that the predictions of the model are not always *instance-wise* accurate (cf. the discussion of Accuracy above). The results in (14a) illustrate a strong preference for *ma* with a probability estimate of *P=0.791*, though in this case the topic marker actually selected is *ne*, which received an estimated *P=0.148*. A possibly relevant factor in this case, though not one of the factors that were incorporated into our model, is the demographic profile of both speakers in file CONV002. Both speakers are college/university educated (in Mandarin) and in their 50's at the time of the recording. Their use of *ne*, in preference to *ma*, a topic marker more associated with informal conversational style, might be a reflection of a slightly more formal style preferred by these speakers.

(14) a. Probability estimates and context for case #40

| | |
|---|---|
| $P(ne \mid C_{\#40})=0.148$ (selected)<br>$P(a \mid C_{\#40})=0.004$<br>$P(mo \mid C_{\#40})=0.049$<br>$P(zi \mid C_{\#40})=0.008$<br>$P(ma \mid C_{\#40})=$**0.791 (predicted)** | Context: topic length 4, clause topic, contrastive function, clause comment-type, conversation genre |

**b.**

| | | | | | |
|---|---|---|---|---|---|
| 现在 | 跑脱 | **呢**， | 也 | 吃 | 个 |
| *ɦi*$^{23}$*zə*$^{53}$ | *bɔ*$^{23}$*tʰə*$^{ʔ55}$ | *ne*$^{ʔ}$， | *ɦa*$^{23}$ | *tɕyə*$^{ʔ55}$ | *gə*$^{ʔ12}$ |
| now | run away | TM | also | eat | SFP |

'Even now when (mother) has passed away, (we) still dine out (once a year).'
(Conv002)

Lastly, in (15), we can observe a case in which all five topic markers are estimated to have approximately equal probability with respect to the observable context, with the estimated probabilities ranging from *P=0.135* for *ne* to *P=0.273* for *mo*. Such instances with close-to-equal estimated probabilities of occurrences could be considered as prime candidates of “true” synonymy, with full interchangeability in the context for the entire selected set of five topic markers. Although, in fact, it was *ne* that was selected by the speaker, the contextual factors are such that the model assigns no really clear winner ahead of the rest of the field. The values for the factors in this case amount to a “messy” scenario: *zi* is preferred with TOPIC-LENGTH of 2 and a VERBAL TOPIC-PART-OF-SPEECH, while *mo* is dispreferred in these same contexts. The model calculates *mo* as having the highest probability in this case, but it is a close contest, especially with *zi* which ends up with the second highest probability. Our Shanghainese co-author’s intuition was that all five topic markers would be viable in (15b).

(15) a. Probability estimates and context for case #94

| | |
|---|---|
| $P(ne \mid C_{\#94})=0.135$ (selected)<br>$P(a \mid C_{\#94})=0.163$<br>$P(mo \mid C_{\#94})=$**0.273 (predicted)**<br>$P(zi \mid C_{\#94})=0.232$<br>$P(ma \mid C_{\#94})=0.197$ | Context: topic length 2, verb topic, introductory function, clause comment-type, interview genre |

b.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 游泳 | 呢, | 我 | 老早 | 老 | 欢喜 | 游泳 | 厄 |
| ɦiɤ$^{23}$ɦioŋ$^{23}$ | ne$^{ʔ}$, | ŋu$^{23}$ | lɔ$^{23}$tsɔ$^{34}$ | lɔ$^{23}$ | hø$^{53}$ɕi$^{34}$ | ɦiɤ$^{23}$ɦioŋ$^{23}$ | ɑ$^{ʔ}$ |
| swim | TM | I | long ago | very | like | swim | SFP |

‘It is swimming that was my favourite sport long ago.’ (Inter005)

In all these results, we see that occurrences of the topic-markers in particular contexts are not categorically determined but are, rather, probabilistic. Furthermore, the contextual variables can only account for the occurrences of the topic-markers in a limited way and there are cases where the model alone can not account for the particular selection that was made by a speaker.

## 7. General discussion

The question of which topic marker in Shanghainese should be used and under what circumstances is not easy to settle. Even when we restrict ourselves, as we have done here, to the five most frequent topic markers in a corpus, there is no simple basis for the selection of a topic marker. To take just one category of a relevant factor – the semantic/pragmatic function of the topic marker – we find that each of the five levels of this factor that we considered (INTRODUCTORY, EMPHATIC etc.) occurs with each of the five topic markers. Similarly, for most of the other factors considered, we find each level of each factor attested for each topic marker. This is a situation where it is simply not possible for an analyst to draw convincing conclusions without the support of statistical analysis. By adopting a multivariate approach to our data, we are able to arrive at an appreciation of the differing extent to which multiple factors, alone or together, play a part in the selection of topic markers. We could report many individual findings concerning this or that category of some variable, the preferences of a topic marker for this or that category of a variable, how particular variables interact etc. There is a danger of losing sight of the larger tendencies in this way of proceeding and it becomes important, therefore, to highlight the key findings emerging from the results.

As illustrated above, a univariate analysis can lead to an appreciation of the factors that favour one topic marker over others and the differing degrees to which each factor plays a part. The differentiation into the three-way classifications of +, 0, - in the summary of the standardized Pearson residuals is one way – and appropriate as the initial way – of gaining some appreciation of the preferences that topic markers show for each category of each factor. In the present study, though, we have focused on a multivariate analysis. This method leads to odds for predicting any one topic marker for each factor level, as in the summary in (10). In addition, the polytomous model allows us to explore the predictions made by the aggregate effect of the coefficients for a combination of features. Determining the strongest predictions for features combined, in turn, leads to identification of best candidates for prototypical usages for each topic marker, as illustrated in Section 6.

We draw attention to some selected findings, without any attempt to summarize each and every result from above. Of particular interest is the behavior of *ma*. A number of results point to *ma* as being a kind of default topic marker in Shanghainese: *ma* shows the least number of significant odds values (just three) in the summary of results from the regression modeling in (10); *ma* shows both the least Accuracy (34%) and the least standard deviation in prediction rates among all topic markers, as mentioned in Section 5. These findings point to *ma* as a relatively general-purpose topic marker, the topic marker that, on the whole, is hardest to predict. This result accords well with the intuition of our Shanghainese-speaking co-author, who describes his own intuition about *ma* in the following terms: "According to my intuition, it is the most natural Shanghainese-specific topic marker which can appear in any environment (after the topic) and replace any other topic markers (with whatever functions they have)".

As follows from the preceding comments on *ma*, the remaining topic markers each have a distinctive profile of preferences and dispreferences in terms of the features they are associated with. Of the more specialized topic markers, *zi* is of particular interest. We know from the historical record that *zi* is the youngest of the five topic markers, having emerged in the Ming Dynasty from earlier copular and focus marking uses (Newman and Han 2013). In both these earlier uses, *zi* functioned to highlight some element(s) on its right, similar to the introductory function 'given-new' function in which the new, more salient information is in the comment structure to the right of the topic marker. These origins of *zi* can still be detected in some of the features preferentially associated with *zi* as a topic marker: the INTRODUCTORY (= 'given-new') FUNCTION and noun as the TOPIC-PART-OF-SPEECH (implying also short topics). The oldest of the topic markers, *ne*, on the other hand, has now come to be strongly associated in Shanghainese with the specialized features relating to the FUNCTION variable: CONTRASTIVE, COUNTEREXPECTED, CONDITIONAL FUNCTIONs, and to a lesser, but still significant degree, EMPHATIC FUNCTION.

The variables that we identified as potentially influencing the choice of topic marker interact with each other in such a way that no one variable categorically determines one and only one topic marker. On the contrary, the variables, taken together, predict outcomes to varying degrees. We are able to arrive at predictions for any of the combinations of variables in the data, but these predictions estimate probabilities for the selection of each topic marker, never a categorical prediction of one and only one topic marker. There can be rather different profiles of estimated probabilities associated with a combination of variables and we have tried to convey some sense of this range through examples (12), (13), (14) and (15).

## 8. Conclusion

We have shown how the suite of functions made available in the `polytomous` package of R provide attractive analytical tools for corpus linguistics attempting to better understand the conditioning of multiple (>2) alternatives. The kinds of multiple alternatives that lend themselves to analysis in these terms are varied and could relate to phonological, lexical or grammatical phenomena. The possibility of many alternative topic markers in a topic-prominent language such as Shanghainese, and the absence of any strict, categorical outcomes in the choice of topic markers, makes the `polytomous` package a natural toolkit to turn to in attempting to make sense of the quite complex data. The final summary of results in (10) and the individual probability estimates for each context type, on the other hand, allow us to make an extremely fine-grained differentiation in the probabilities associated with each combination of values of contextual factors. Taken altogether, this way of proceeding offers, we think, a highly satisfying account of the sometimes quite subtle factors underlying the choice of topic marker in Shanghainese. With respect to other types of polytomous linguistic alternations, one can mention studies on synonymy (Arppe 2008, 2009: Finnish THINK verbs; Divjak and Arppe 2013: Russian TRY and Finnish THINK verbs), allophonic variation (Arppe and Tucker 2012: English /t/ allophones), and constructional, or syntactic alternations (Arppe 2011: English ACTIVE vs. *be/get/become* PASSIVEs).

We have framed our study in terms of moving from observation to prediction (as reflected in the title), aided by the `polytomous` package. But it is natural and desirable to move beyond the corpus-based study to more experimental studies that seek to determine the degree of psychological reality associated with the findings of our study. For the cases examined individually in Section 6, the corpus-based model produced probability estimates that did seem overall in accord with the native speaker perceptions of our Shanghainese-speaking co-author, agreeing that a number of alternatives seemed equivalent or that one alternative seemed the best, etc. Such native speaker responses are somewhat reassuring and inspire confidence that psycholinguistic experimental work may well provide confirmation of the psychological realities of the results of the model. The estimated probabilities of topic markers that we obtained for each context type offer an ideal starting point for follow-up psycholinguistic studies. Forced-choice experiments, where speakers are forced to choose between topic markers in particular contexts, suggest themselves as one experimental approach to take, with the possibility of comparing results from such experiments with the estimated probabilities derived from our corpus-based study. Preliminary results from psycholinguistic studies of this type, as a way of confirming predictions from the polytomous regression analysis, are encouraging. Ultimately, it is through a multi-methodological approach, rather than an approach based on any one method, that a full understanding of Shanghainese topic marking will emerge.

## Abbreviations

TM = topic marker, SFP = sentence-final particle

## Acknowledgements

The work of the first author was supported by a grant from the China Institute, University of Alberta. The work of the second author was supported by a postdoctoral fellowship from the University of Helsinki as well as a grant for research collaboration abroad from the Academy of Finland (#136322/2009). The authors are grateful to two anonymous reviewers and to the editor of the journal for extensive and helpful feedback on an earlier draft of this article.

## Bionote

Weifeng Han is a Lecturer in the Department of English, Donghua University, Shanghai. His research interests include cognitive linguistics, syntactic typology, linguistic philosophy and second language acquisition.

Antti Arppe is an Assistant Professor in Quantitative Linguistics in the Department of Linguistics, University of Alberta. His research interests include corpus linguistics, specifically exploiting and developing statistical methods, and in general multimethodological, empirical research strategies in linguistics, and the study of various sorts of linguistic alternations.

John Newman is a Professor in the Department of Linguistics, University of Alberta. His research interests include corpus linguistics, cognitive linguistics, and field work (in Papua New Guinea). He is the Director of ICE-CANADA, the Canadian component of the International Corpus of English.

Notes

[1] In some formal analyses (cf. Gasde and Paul 1996), the topic marker is viewed as the head of a topic structure. Han (2010), however, proposes that the head of a topic structure is the topicalized content rather than the topic marker. The issue of which element should be considered the head of the topic structure is not germane to the present discussion.

[2] A happy and unexpected result from our corpus-based approach has been the identification of 12 Shanghainese topic markers, hitherto overlooked in previous literature on the subject: *mə*$^{ʔ12}$ (嚜), *ma* (嘛), *gə*$^{ʔ12}$*ɦɛ*$^{23}$*ɦo*$^{23}$ (个闲话), *ɦɛ*$^{23}$*ɦo*$^{23}$ (闲话), *la*$^{23}$ (啦), *tɕiɔ*$^{34}$ (叫), *fɛ*$^{34}$*tɔ*$^{34}$*zɿ* (反倒是), *gə*$^{ʔ12}$*ɦɛ*$^{23}$*ɦo*$^{23}$*mə*$^{ʔ}$ (个闲话末), *ɦɛ*$^{23}$*ɦo*$^{23}$*ɲi* (闲话呢), *mə*$^{ʔ}$*ɲi* (末呢), *ɲizɿ*$^{23}$ (呢是) and *mə*$^{ʔ}$*zɿ*$^{23}$ (末是).

[3] It has not been usual to consider a variable defined in terms of the number of syllables in the topic structure when analyzing topic-comment structures. But we chose to include this variable as a new possibility worth exploring.

[4] We use FINALTAG to cover the tag of Shanghainese tag questions, such as *tɛ*$^{34}$*va*$^{ʔ12}$ (对伐), and other utterance-final discourse particles which play a part in signalling turn-taking, such as the *zɿ*$^{23}$*va*$^{ʔ12}$ (是伐) in example (3).

[5] In our general experience working with polytomous logistic regression modeling of various linguistic phenomena in a number of languages, $R_L^2$ values approaching 0.3 can be considered quite good, and it is difficult to push this performance value beyond 0.4 without overfitting the model. Nevertheless, for the sake of comparison, we fitted a support-vector machine (SVM) with the same data and explanatory variables, reaching an *Accuracy* of 0.516 and a $R_L^2$ of 0.247. As can be noted, the performance of the two different methods are very close to each other.

[6] Since the constituent binary models are fit separately of each other, their instance-wise probability estimates do not necessarily exactly sum up to the theoretically correct $\sum_{Topic\text{-}Marker} P(Topic\text{-}Marker|Context)=1.0$. Consequently, the probability estimates are adjusted so that $\sum P=1.0$ by simply dividing instance-wise each original topic-marker-specific probability estimate by the sum of these estimates for that particular instance.

**References**

Agresti, Alan. 2002. *Categorical Data Analysis*, 2$^{nd}$ edn. Hoboken: John Wiley & Sons.

Arppe, Antti. 2008. *Univariate, Bivariate and Multivariate Methods in Corpus-based Lexicography - a Study of Synonymy*. Helsinki, Finland: University of Helsinki dissertation.

Arppe, Antti. 2009. Linguistic choices vs. probabilities – how much and what can linguistic theory explain? In Sam Featherston and Susanne Winkler (eds.), *The Fruits of Empirical Linguistics* (Volume 1: Process), 1-24. Berlin: Mouton de Gruyter.

Arppe, Antti. 2011. From modeling lexical synonyms to constructional alternations. *Proceedings of the Corpus Linguistics Conference 2011* (CL2011), University of Birmingham, England, 20-22 July 2011.

Arppe, Antti. 2013. *Package 'polytomous': Polytomous Logistic Regression for Fixed and Mixed Effects* (Version 0.1.6). The R Project for Statistical Computing. http://cran.r-project.org/web/packages/polytomous/index.html

Arppe, Antti, Weifeng Han, and John Newman. in prep. Topic marking in a Shanghainese corpus: what simple univariate and bivariate methods can tell.

Arppe, Antti and Juhani Järvikivi. 2007. Every method counts: Combining corpus-based and experimental evidence in the study of synonymy. *Corpus Linguistics and Linguistic Theory* 3(2). 131-159.

Arppe, Antti and Benjamin V. Tucker. 2012. You should model what you observe – the case of the allophonic realizations of the English /t/ in a spontaneous speech corpus. *Pre-proceedings of the International Conference on Linguistic Evidence*. Tübingen, Germany, 9-11 February 2012.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina and R. Harald Baayen. 2007. Predicting the Dative Alternation. In Gerlof Boume, Irene Krämer and Joost Zwarts (eds.), *Cognitive Foundations of Interpretation*, 69-94. Amsterdam: Royal Netherlands Academy of Science.

Chafe, Wallace. 1976. Givenness, contrastiveness, definiteness, subjects, topics and point of view. In Charles N. Li (eds.), *Subject and Topic*, 25-55. New York: Academic Press.

Chu, Chauncey C. 1987. *Historical Syntax-Theory and Application to Chinese*. Taipei: The Crane Publishing.

Copeland, James E. and Philip W. Davis. 1983. Discourse portmanteaus and the German Satzfeld. In William Agard, Gerald Kelly, Adam Makkai and Valerie Makkai (eds.), *Essays in Honor of Charles F. Hockett*, 214-245. Leiden: Brill.

Divjak, Dagmar and Antti Arppe. 2013. Extracting prototypes from exemplars. What can corpus data tell us about concept representation? *Cognitive Linguistics* 24(2): 221-274

*Encyclopedia of Shanghai*. 2010. *Shanghai Baikequanshu* [The Encyclopedia of Shanghai]. Shanghai: Shanghai Scientific and Technical Publishers.

Featherston, Sam. 2005. The decathlon model. In Stephan Kepser and Marga Reis (eds.), *Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives*, 187-208. Berlin & New York: Mouton de Gruyter.

Frank, Eibe and Stefan Kramer. 2004. Ensembles of nested dichotomies for multiclass problems. In Carla E. Brodley (eds.), *Proceedings of the 21st International Conference on Machine Learning*, 305-312. ACM Press.

Gasde, Horst-Dirter and Waltraud Paul. 1996. Functional categories, topic prominence and complex sentences in Mandarin Chinese. *Linguistics* 34(2). 263-294.

Gries, Stefan Th. 2003. Towards a corpus-based identification of prototypical instances of constructions. *Annual Review of Cognitive Linguistics* 1(1). 1-27.

Gundel, Jeanette K. 1985. Shared knowledge and topicality. *Journal of Pragmatics* 9(1). 83-107.

Halliday, M. A. K. 1985. *An Introduction to Functional Grammar*. London: Edward Arnold.

Han, Weifeng. 2010. *A Typological Study of the Syntactic Properties of Topic and Topic Markers*. Shanghai: Shanghai International Studies University dissertation.

Hanks, Patrick. 1996. Contextual dependency and lexical sets. *International Journal of Corpus Linguistics* 1(1). 75-98.

Harrell, Frank E. 2001. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer-Verlag.

Hosmer, David W. Jr. and Stanley Lemeshow. 2000. *Applied Regression Analysis*. New York: Wiley.

Kuno, Susumu.1973. *The Structure of the Japanese Language.* Cambridge Mass: MIT Press.

Lambrecht, Knud. 1988. Presentational cleft constructions in spoken French. In John Haiman and Sandra A. Thompson (eds.), *Clause Combining in Grammar and Discourse*, 135-179. Amsterdam: John Benjamins.

Li, Charles N. (eds.). 1976. *Subject and Topic.* New York: Academic Press.

Li, Charles N. and Sandra A. Thompson. 1976. Subject and topic: A new typology of language. In Charles N. Li (eds.), *Subject and Topic*, 457-489. New York: Academic Press.

Li, Liqun. 2010. Discussion of the topic marker *wo shuo*. *Journal of Hexi College* 26(3). 96-99.

Li, Yanyan. 2010. On the topic marker *yaoshuo*. *Journal of Tangshan Teachers College* 32(1). 25-27.

Liu, Danqing. 2004. Identical topics: A more characteristic property of topic-prominent languages. *Journal of Chinese Linguistics* 32(1). 20-64.

Menard, Scott. 1995. Applied logistic regression analysis. In Michael S. Lewis-Beck (eds.), *Sage University Paper Series: Quantitative Applications in the Social Sciences*, 07-106. Thousand Oaks, CA: Sage Publications.

Newman, John and Weifeng Han. 2013. The topic marker zi (是) in contemporary Shanghainese dialect. Presentation to the American Association of Corpus Linguistics, San Diego State University, San Diego, USA.

Qian, Nairong, Baohua Xu and Zhenzhu Tang. 2007. *Shanghai hua da cidian* [A Comprehensive Dictionary of Shanghainese]. Shanghai: Shanghai Cishu Publishing House.

R Development Core Team. 2012. *R: A Language and Environment for Statistical Computing*. http://www.R-project.org.

Reinhart, Tanya. 1981. Pragmatics and linguistics: An analysis of sentence topic. *Philosophica* 27(1). 53-94.

Rifkin, Ryan and Aldebaro Klautau. 2004. In defense of one-vs-all classification. *Journal of Machine Learning Research* 5 (Jan). 101-141.

Schiffrin, Deborah. 1992. Conditionals as topics in discourse. *Linguistics* 30(1). 165-197.

Tomlin, Russell S. 1985. Foreground-background information and the syntax of subordination. *Text* 5(1-2). 85-122.

Xu, Jingyi. 2010. A study on the topic markers of *guanyu*, *duiyu* and *zhiyu*. Shanghai: Shanghai Normal University MA thesis.

Xu, Liejiong and Danqing Liu. 1998. The copied topic structure in Mandarin and Shanghainese. *Language Teaching and Study* (20)1. 85-103.

Xu, Liejiong and Danqing Liu. 2007 [1998]. *Topic: Structural and Functional Analysis,* 2nd edn. Shanghai: Shanghai Education Publishing House.