Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 259–267
Marseille, 11–16 May 2020

Do You Believe It Happened? Assessing Chinese Readers' Veridicality Judgments

Yu-Yun Chang[†], Shu-Kai Hsieh[‡]
Graduate Institute of Linguistics, National Chengchi University[†],
Graduate Institute of Linguistics, National Taiwan University[‡]
No. 64, Sec. 2, ZhiNan Rd., Wenshan District, Taipei City 11605, Taiwan (R.O.C)[†],
No. 1, Sec. 4, Roosevelt Rd., Taipei City 10617, Taiwan (R.O.C.)[‡]
yuyun@nccu.edu.tw, shukaihsieh@ntu.edu.tw

Abstract

This work collects and studies Chinese readers' veridicality judgments to news events (whether an event is viewed as happening or not). For instance, in "The FBI alleged in court documents that Zazi had admitted having a handwritten recipe for explosives on his computer", do people believe that Zazi had a handwritten recipe for explosives? The goal is to observe the pragmatic behaviors of linguistic features under context which affects readers in making veridicality judgments. Exploring from the datasets, it is found that features such as event-selecting predicates (ESP), modality markers, adverbs, temporal information, and statistics have an impact on readers' veridicality judgments. We further investigated that modality markers with high certainty do not necessarily trigger readers to have high confidence in believing an event happened. Additionally, the source of information introduced by an ESP presents low effects to veridicality judgments, even when an event is attributed to an authority (e.g. "The FBI"). A corpus annotated with Chinese readers' veridicality judgments is released as the Chinese PragBank for further analysis.

Keywords: event, veridicality, pragmatics, linguistic features, readers' commitments

1. Introduction

While reading news, for instance, "The FBI alleged in court documents that Zazi had admitted having a handwritten recipe for explosives on his computer", do people believe that *Zazi had a handwritten recipe for explosives*? On the other hand, what do people infer if the sentence is "According to the FBI agents, there is relatively little evidence that Zazi had a handwritten recipe for explosives"? Do they still believe in the event? This kind of judgments in believing whether an event described in a sentence happened or not is called veridicality. The term, veridicality defined by Giannakidou (1999), regards the degree of commitment as a gradable distribution among categories. de Marneffe et al. (2012) suggest using the term veridicality to depict readers' degree of commitments to an event. Therefore, veridicality is different from factuality used in Saurí (2008)'s study, which assess speakers' degree of commitments to an event. This study focuses on assessing readers' veridicality judgments to news events at sentence level.

For readers, in order to believe whether an illustrated event happened or not, they will need to assess the information conveyed by the speaker to some extent as well (de Marneffe et al., 2012). This also infers readers may not hold the same commitments to an event as the speakers. Nowadays, the news events published online can be broadly read by thousands of people. Therefore, it is essential to study how a large number of readers interpret the events, and to explore the factors that influence readers in making veridicality judgments.

There are some veridicality-related corpora available, such as MEANTIME corpus (van Son et al., 2014), an extension of Ita-TimeBank (Minard et al., 2014), and FactBank (Saurí, 2008; Saurí and Pustejovsky, 2009). However, all of these collect judgments from the speakers' perspectives instead of readers'. Based on FactBank corpus, de Marneffe et al. (2012) built the English PragBank (a total of 642 sentences selected from FactBank), which collected readers' veridicality judgments to news events. The English PragBank takes the encyclopedic knowledge into account while collecting judgments; whereas, the FactBank studies factuality judgments mainly based on lexical theory.

For example, while considering the word *say* under a lexical perspective, it can only be analyzed as *non-veridical* (which means that the word *say* does not imply that the illustrated event is a fact (true) or counterfact (false) in the real world). However, as the news event listed in example (1) retrieved from de Marneffe et al. (2012)'s paper, the word *say* marks up the source of information *United Widget*, which affect readers to have higher reliability in believing that the chairman of United Widget resigned. As shown in the above cases, readers' veridicality judgments should not only be assessed under lexical concerns, but under a pragmatic viewpoint based on the context of the event sentence.

(1) United Widget said that its chairman resigned.

However, up to now, there is no corpus collecting veridicality judgments from readers' perspectives in Chinese. Therefore, this study collects datasets annotated with veridicality judgments of readers in Taiwan (the Chinese PragBank), and explores linguistic features embedded with contextual factors that affect readers' veridicality judgments. This study aims at a

better understanding and characterization of the context in which events are embedded and how the context leads to human judgments of event veridicality. The goal of this study is to explore and analyze the pragmatic behaviors of linguistic cues derived from theories systematically, in order to help machine learning models to predict readers' veridicality judgments in the near future application.

2. Related Work

2.1. Modality and Evidentiality

Since readers' veridicality judgments may be affected by speakers' commitments to an event, the linguistic notions considered in factuality are taken into account in this study as well, such as modality and evidentiality. Modality markers like *certain*, *probable* and *possible* convey different degrees of possibilities. Some researchers had divided modality into two main categories (epistemic modality and deontic modality) (von Wright, 1951; Lyons, 1977; Steele et al., 1981), and others proposed grouping modality into three categories (epistemic modality, deontic modality, and dynamic modality) (Palmer, 2001; Tsang, 1981; Tiee, 1985; Hwang, 1999). Among the three modalities, epistemic modality has an impact on factuality (Saurí and Pustejovsky, 2009), which expresses the speaker's opinions of the truth value within a proposition. Thus, epistemic modality is considered while assessing veridicality. Although in Chinese, Huang et al. (2017) proposed a framework for differentiating Chinese modalities, the lexical forms of modalities are still open to debate. The above approaches only considered grouping modalities under grammatical notion. Hsieh (2005) suggested that Chinese modalities should also be analyzed semantically. She elaborated that the occurrence of Chinese modality in a sentence denoted an implicit semantic source, which refers to speaker's judgment, opinion or attitude to the event. This semantic perspective of Chinese modalities is taken into account while exploring readers' veridicality judgments.

As for evidentiality, since it is defined as the way how source of information is acquired (van Valin and LaPolla, 1997; Aikhenvald, 2004; Jakobson, 1957), it will show an impact on veridicality judgments. Mushin (2001) also addressed that evidentiality can not only specify the source of information, but speaker's epistemic attitude to the event. In Chinese, the issues of the definition of evidentiality or how to classify evidentiality are still controversial (Zhu, 2006; Ma, 2011; Su and Liu, 2012). It is found that evidential markers embedded with various degrees of commitments can be expressed via different types of predicates, such as predicates expressing opinion or belief (e.g., *think*, *suspect*), attempt (e.g., *attempt*, *try*), and command (e.g., *call for*, *order*) (Saurí, 2008). These predicates are also known as event-selecting predicates (ESPs) (Saurí and Pustejovsky, 2009). The ESP is a kind of predicate which selects an event as its argument. As shown in example (2) (Saurí, 2008), the ESP "suspects" selects an event "Freidin left the country in June" as its argument. The identification of an event follows the guidelines in TimeBank (Pustejovsky et al., 2003a; Pustejovsky et al., 2003b), specifying an event is usually expressed via a predicate.[1]

(2) Berven *suspects* that Freidin **left** the country in June.

In addition, an ESP can also introduce the source of information to an event. For instance in the above example, the source of information "Berven" is introduced by the ESP "suspects", which expresses that Berven presents it is possible "Freidin left the country in June". As shown in (1), veridicality judgments may be affected by the source of information of an event as well. Therefore, in this study, we focus on exploring event sentences with source of information introduced by ESPs.

It is noted that readers' veridicality judgments should not only be assessed on linguistic theories, but also on the effects brought by readers' encyclopedic knowledge. Thus, this is also a kind of subjectivity study (Saurí and Pustejovsky, 2009), which involves a person's psychological viewpoints (Banfield, 1982; Wiebe, 1994). Namely, it involves private states of a person's mind, which includes thoughts, emotions, perceptions and attitudes. In addition, these states, as noted by Quirk et al. (1985), could not be observed or verified through objectivity. Therefore, the study of readers' veridicality judgments considers linguistic notions under pragmatic usage in real world application, including personal perspectives and subjective interpretations to an event.

2.2. The Scalability of Veridicality

To scale the degree of commitment to an event, researchers have suggested different approaches. Rubin et al. (2006) proposed a Four-Dimensional Certainty Categorization Model for a certainty identification task. In the model, the first dimension presented that the certainty level could be divided into four hierarchical categories, which were absolute, high, moderate, and low. This model has been further adopted into the research by Su et al. (2010), which applied the first dimension to the categorization of evidentiality in the detection of text trustworthiness on Collaborative Question Answering. However, within the first dimension, the definition of the moderate category is vague. For example while assessing readers' veridicality judgments, a moderate value could have two different interpretations: 1) a reader might have 50% chance in believing the event happened or not; or 2) a reader does not have specific preference or interests on that event which denotes an unknown category. It would be hard to learn which interpretation was indicated by the reader. Therefore, these four categories of certainty level are not able to capture veridicality adequately.

Another scale used in assessing degree of commitments was introduced by Lee et al. (2015), which proposed

[1] In this paper, an ESP is italicized and an event is boldfaced in all given examples.

inviting non-expert workers in identifying speakers' commitments to events, based on a scale of -3 (certainly did not happen) to 3 (certainly did happen), where 0 was included and denoted that the speaker was neutral and presented no bias to the commitment of the event. Since the 0 value was presented on a scale, it would fall into the two-interpretation situation as the moderate category mentioned above. Therefore, the approach proposed by Lee et al. (2015) may not be an appropriate scale for assessing readers' veridicality judgments.

Apart from the above frameworks in categorizing or scaling degree of commitments, Saurí and Pustejovsky (2009) proposed a scale with different factuality categories based on Horn (1989)'s study, which both logical and linguistic aspects were taken into account. The scale is composed of polarities (depicting whether an event happened or not) and probabilities (expressing degree of certainty, which are certain, probable, and possible). Seven of the factuality categories are adopted by the English PragBank to assess readers' veridicality judgments, which are certainly happened (`CT+`), probably happened (`PR+`), possibly happened (`PS+`), certainly not happened (`CT-`), probably not happened (`PR-`), possibly not happened (`PS-`), and unknown (`Uu`). This scale of seven veridicality categories is also applied in this study.

3. Data Collection

The news sentences in this study are crawled from PTT in Taiwan. PTT is a bulletin board system that is prevalently used in Taiwan. A number of various news posts are posted on its Gossiping Board with the tag 新聞 'news' attached to the titles everyday.[2] A total of 968 news sentences are extracted.

In this study, each event is identified by locating an ESP in a sentence. All the veridicality judgments are annotated by annotators who are non-expert in linguistics to reveal their most straight forward and direct veridicality judgments to each event, without being intervened by linguistic knowledge. Each item contains an event sentence, a normalized sentence and a scale with seven veridicality categories. A normalized sentence removes all the negation markers (e.g., 不 'not') and modality markers (e.g., 可能 'possible') from the target event for annotators to easily focus on the target event, as conducted in Saurí and Pustejovsky (2009) and de Marneffe et al. (2012)'s studies. Examples of normalized sentences are presented below, (3) and (4).

(3) 顯鈞 *坦言* 在校 成績 不 **佳**
Xian-Chun admit in school grade not good
'Xian-Chun *admitted* that he doesn't have **good** grades in school.'

Normalization: 顯鈞在校成績佳 'Xian-Chun has good grades in school.'

(4) 有 消息 *指出* ，金元弘
there is news indicate Kim Won Hong
可能 是 涉及 **貪污**
possible is engage corruption
濫權 遭 拔官。
abuse of authority be dismiss
'There is news *indicates* that Kim Won Hong was dismissed from his position for possible engagement in **corruption** and abuse of authority. '

Normalization: 金元弘貪污 'Kim Won Hong engaged in corruption.'

The items present to each annotator are shuffled. After reading the event sentence, annotators are asked to assign a veridicality judgment from the scale based on the normalized sentence. The veridicality scale on the questionnaire is translated into Chinese.

This study follows the annotation guidelines as proposed in the English PragBank, and collects two datasets. The first dataset is displayed via Qualtrics platform, and a total of 288 annotators (aged 18-33) are recruited by crowd-sourcing and are given with a compensation of NT$250 per task. Each task is presented in blocks of 45 target sentences and 5 filter sentences (three positive and two negative sentences). The filter sentences are non-corpus sentences with correct answers, which are taken to identify whether annotators are paying attention to the task or are behaving as outliers. If an annotator does not answer one of the five filter sentences correctly, the responses collected from the annotator are removed from the dataset. The answers to the filter sentences are annotated and fully agreed by the other 10 people who are not participants of this experiment. An example of negative filter sentence is shown in (5). A total of 151 annotators (around 3/4 of the 208 annotators) answered the filter sentences correctly. Through this approach, each item is annotated with at least 6 judgments.

(5) 北投 一棟 廢棄 空屋
Beitou one abandoned empty house
昨 半夜 發生 火警 所幸
yesterday midnight happen fire fortunately
無 人 **傷亡**
no people casualty
'Yesterday at midnight, fire broke out in an abandoned empty house in Beitou, and fortunately there were no **casualties**.'

Normalization: 有人傷亡 'There were casualties.'

Annotations: `CT-`:10

Since this study asks readers to make their judgments straightforwardly, we would like to know whether a dataset without filter sentences will have a big difference as compared to the first dataset. Therefore, a second dataset without filter sentences annotated with

[2] In this study, the earliest crawled news article is posted on Feb. 26, 2017.

readers' veridicality judgments is collected for comparison. In the second dataset, each annotator annotated 200 items via Google Form. There are 35 out of 48 annotators completed the task, aged 26-30. In this approach, each item is annotated with at least 6 judgments as well.

To be simple, the dataset collected from the first approach is called the Qualtrics dataset; and the one gathered from the second approach is the Google dataset.

4. Analysis of Readers' Veridicality Judgments

Figure 1 shows a summary of distribution types among veridicality judgments of the two datasets. The labels on x-axis show the number of sentences and y-axis represent types of distributions. For example, group 3/3 indicates sentences in the group are evenly annotated with two veridicality categories (e.g., a sentence which 3 participants assigned `CT+` and the others assigned `PR+` to the event; or a sentence which 3 participants annotated as `CT+` and others annotated as `PS+`). Group 6 presents sentences that all annotators have in agreement.

In general, the types of distributions between the two datasets presented in the figure are quite similar. As seen in Figure 1, most sentences are tagged into group 1/2/3 in both datasets, and a large amount of sentences are not annotated with more than half of the same veridicality judgments (e.g., group 3/3, 1/2/3, 1/1/1/3, 2/2/2/, 1/1/2/2, 1/1/1/1/2, and 1/1/1/1/1/1). Further details of the datasets are presented below.

4.1. Inter-annotator Agreement

In order to examine the inter-annotator agreements of the two datasets, five statistical measurements of reliability coefficients among annotators are applied, which are Fleiss's kappa (Fleiss, 1981), Krippendorf's α (Krippendorff, 1980), Intraclass correlation (ICC) (Shrout and Fleiss, 1979), Robinson's A (Robinson, 1957), and Finn's r (Finn, 1970). Fleiss's kappa, which is an extension of Cohen's kappa, has the advantages that it is able to measure the reliability of multiple annotators, and different items can be rated by different people. However, it is a conservative measure and deals with nominal categories which does not take the order of the 7 veridicality categories into consideration. Thus, the other four measurements are introduced which view the veridicality scale as a continuum. The advantages of applying the four measurements are briefly stated in the following. For Krippendorf's α, it deals with any types of datasets, and calculates disagreements between raters rather than agreements; and for ICC, which is an improvement of Spearman's rho, takes the variance into consideration, and calculates consistency (irrelevant consistency) and agreement (relevant consistency). As for Robinson's A, it calculates disagreements among raters as well, and is suggested to be used when internal consistency of the raters is poor; whereas, Finn's r assumes equal distribution in all categories, and is suggested to be used when agreement among raters is high. Below Table 1 shows the results of inter-annotators' reliability coefficients calculated by the five measurements.

Measurements	**Qualtrics dataset**	**Google dataset**
Fleiss's kappa	0.20	0.12
Krippendorf's α	0.30	0.20
ICC consistency	0.30	0.20
ICC agreement	0.30	0.20
Robinson's A	0.42	0.34
Finn's r	0.40	0.28

Table 1: The inter-annotators' agreements of the Qualtrics and Google datasets

As seen from the results listed in Table 1, the overall agreements of the two datasets are quite low. For example, usually the threshold of a satisfactory kappa value is set at 0.6; however, the Fleiss's kappa values in the Qualtrics dataset and Google dataset only present 0.20 and 0.12 respectively. It is also observed that the overall agreements of the Qualtrics dataset are slightly higher than the Google dataset. This indicates that annotators in the Qualtrics dataset have slightly higher confidence in making judgments than in the Google dataset, even for cases containing less informative messages. For example, in (6), even though annotators do not know who 張妻 'Mrs. Chang' is, all the 6 annotators in the Qualtrics dataset fully believe that the event certainly happened; whereas, in the Google dataset, only 3 annotators tag the event as `CT+`. In addition, in (7), the source of information provided by a pronoun 他 'he' is also a less informative message because readers are not able to refer back to know the referee of the speaker in this event. Again in this case, annotators in the Google dataset are less confident in believing the event happened (with 3 people agree on `Uu` label and the others agree on `PR+`, `PS+`, and `PR-`); whereas annotators in the Qualtrics dataset all give high agreements on considering the event as certainly happened (`CT+`).

(6) 張妻 提訟 *表示*，跟
Mrs. Chang file a lawsuit say with
丈夫 在 1996 年 **結婚**
husband in 1996 year marry
'Mrs. Chang filed a lawsuit and *said* that she **married** to her husband in the year of 1996.'

Normalization: 張妻跟丈夫在 1996 年結婚 'Mrs. Chang married to her husband in the year of 1996.'

Qualtrics Annotations: `CT+`:6

Google Annotations: `CT+`:3, `PR+`:2, `Uu`:1

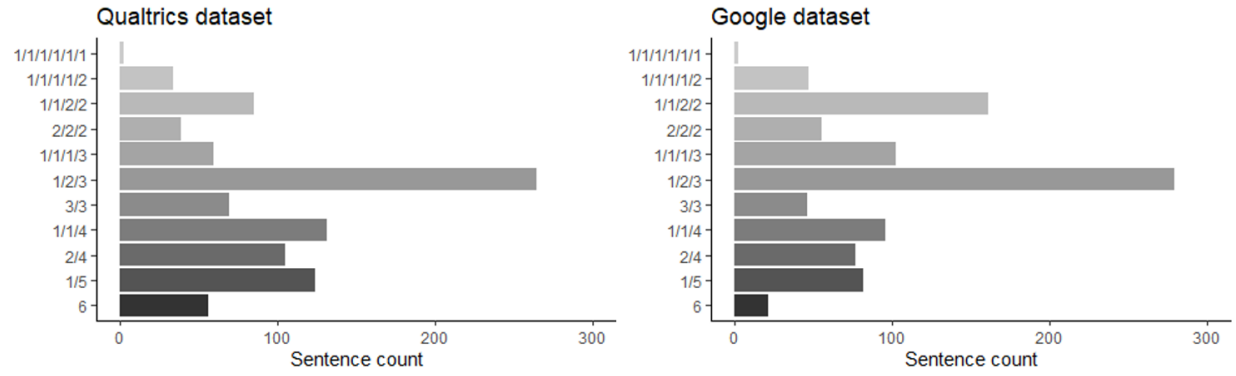

Figure 1: Types of reaction distribution

(7) 他 *認為* **薪資** 短少
he think salary cut
數萬元
several tens of thousands
'He *thinks* that there is a pay cut of several tens of thousands in his **salary**.'

Normalization: 薪資短少數萬元 'There is a pay cut of several tens of thousands in his salary.'

Qualtrics Annotations: CT+:6

Google Annotations: PR+:1, PS+:1, PR−:1, Uu:3

The reason that leads to this effect may due to the design of filter sentences that are added into the Qualtrics dataset. As stated before, if annotators do not answer one of the filter sentences correctly, their judgments in the task will be removed from the dataset. A total of 57 people failed to answer one of the filter sentences correctly. Furthermore, it is observed that 40 out of 57 people (70%) are simply making judgments with different degrees of confidence, rather than making judgments with different polarities. Although the original purpose of applying filter sentences is to filter out annotators that are not being attentive to the task and to increase the inter-annotator agreement, this approach also raises a bar by only allowing annotators who tend to have higher confidence while making judgments to participate in the task. This phenomenon further explains why the Qualtrics dataset has higher inter-annotator agreements than the Google dataset.

4.2. Features with Contextual Factors

In order to observe the linguistic features with contextual behaviors systematically for future application, we focus on analyzing the sentences with a majority vote in both datasets. Sentences with a majority vote are sentences where there are at least 4 annotators making the same veridicality judgments. There are 416 sentences with a majority vote in the Qualtrics dataset, and 276 sentences found in the Google dataset. Figure 2 shows the distribution of which veridicality judgments these sentences are mainly assigned to.

As presented in Figure 2, around 75% of sentences with a majority vote are assigned with a positive polarity in both datasets, which denotes that most annotators believe the events happened/is happening/will happen with different degrees of confidence. With the effect of filter sentences which evokes annotators in the Qualtrics dataset to have higher confidence in making judgments, its number of sentences with a majority vote is 1.5 times higher than the Google dataset.
To find out the shared linguistic features from both datasets in affecting Chinese readers' veridicality judgments, we further investigate the datasets under three perspectives, which are sentences that are assigned with the same majority judgments in both datasets, sentences with a majority vote in the Qualtrics dataset but not in the Google dataset, and sentences with a majority vote in the Google dataset but not in the Qualtrics dataset.

4.2.1. Features with High Confidence

Observed from the datasets, the following features would trigger readers to have higher confidence in believing an event certainly happened or not. Firstly, while an event is introduced by an ESP such as 宣布 'announce', 發布 'post', 認定 'affirm', 查出 'find out', 發表 'publish', 坦承 'admit', 坦言 'admit', 自爆 'say (by oneself)', 發現 'find', 證實 'prove', 顯示 'reveal', 調查 'investigate', 強調 'emphasize', or 承認 'admit', it expresses a statement with high certainty and evoke readers to have high confidence while making judgments. For example, the ESPs 宣布 'announce' in (8) and 坦承 'admit' in (9) evoke at least four annotators to tag the events as CT+.

(8) UBER 台灣 即 *宣布* **暫停**
UBER Taiwan now announce out of service
叫車服務
taxi service
'UBER Taiwan *announces* that the taxi is **out of service** now.'

Normalization: UBER 台灣暫停叫車服務 'The taxi of UBER Taiwan is out of service.'

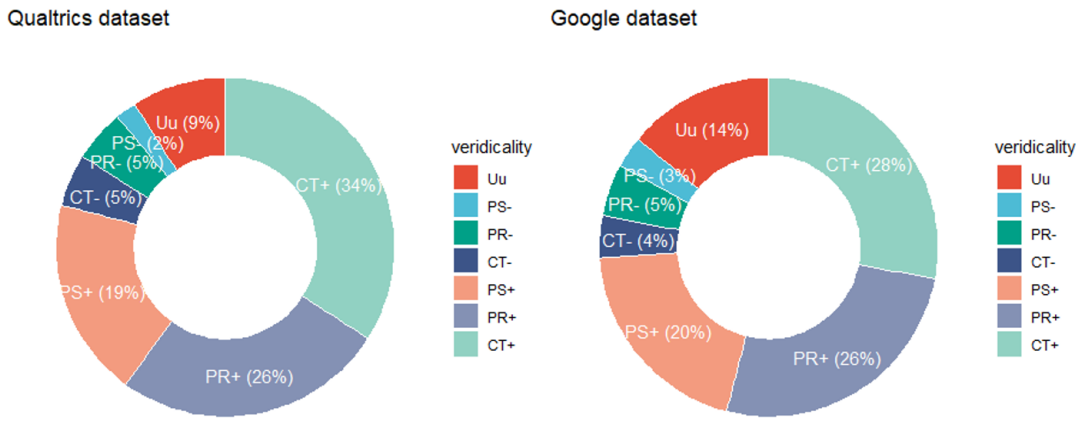

Figure 2: The distribution of sentences with a majority vote

Qualtrics Annotations: CT+:4, PR+:1, PS+:1

Google Annotations: CT+:4, PR+:2

(9) 李男 *坦承* 向 身分不明 男子 **買**
Mr. Li admit from unidentified man buy
安非他命
meth
'Mr. *admitted* **buying** meth from an unidentified man.'

Normalization: 李男向身分不明男子買安非他命 'Mr. Li bought meth from an unidentified man.'

Qualtrics Annotations: CT+:4, PR+:1, Uu:1

Google Annotations: CT+:4, PR+:1, PS+:1

In addition, if an event is described with statistics (e.g., '39 people' in (10)) or temporal information (e.g., 'last year' in (10) or 'last April' in (11)), it would also trigger readers to have higher confidence in believing the event happened.

(10) 市府 衛生局
Taoyuan City Department of Public Health
統計， 桃園市 去年 有 三九
calculate Taoyuan City last year have 39
人 **自殺** 死亡
people suicide dead
'The Department of Public Health in Taoyuan City has calculated that there are 39 people die by suicide in Taoyuan City last year.'

Normalization: 桃園市去年有三九人自殺死亡 'There are 39 people die by suicide in Taoyuan City last year.'

Qualtrics Annotations: CT+:4, PR+:1, PS+:1

Google Annotations: CT+:6

(11) 勞動部 *指出*， 該案 **發生**
Ministry of Labor point out the case happen
在 去年 4 月
in last year April
'The Ministry of Labor pointed out that the case happened last April.'

Normalization: 該案發生在去年 4 月 'The case happened last April.'

Qualtrics Annotations: CT+:5, PR+:1

Google Annotations: CT+:4, PR+:1, PS+:1

4.2.2. Features with Low Confidence

To evoke readers in making probable, possible or unknown judgments to an event, ESPs that are not covered in the above Section 4.2.1. would show an effect, such as 懷疑 'doubt', 控訴 'accuse', 估計 'estimate', 推測 'estimate', 爆 'disclose', 透漏 'disclose', 自稱 'claim (by oneself)', 傳出 'it is heard that', 研判 'infer', 解釋 'explain', 揭露 'disclose', 宣稱 'claim', and 指控 'accuse'. Examples of the ESPs 懷疑 'doubt' and 控訴 'accuse' are listed in (12) and (13) respectively.

(12) 張妻 *懷疑* 兩人 已 有
Mrs. Chang suspect two people already have
多次 **肌膚之親** 的 超友誼，
multiple sex 's intimate relationship
要求 葉女 賠償 50
request Ms. Yeh compensate 50
萬元 精神 慰撫金。
million dollars spiritual compensation
'Mrs. Chang *suspected* that the two people (Ms. Yeh and her husband) had **sex** multiple times, and claimed a compensation of NT$ 500,000 from Ms. Yeh.'

Normalization: 兩人已有多次肌膚之親 'The two people had sex multiple times.'

Qualtrics Annotations: PR+:3, PS+:2, Uu:1

Google Annotations: PR+:2, PS+:1, PS−:1, Uu:2

(13) 2 名 實習 女大生 *控訴* 實習
two intern female undergrad accuse intern
期間 每天 工時 **超過** 14 小時
period everyday working hour over 14 hour
'Two female undergrad interns *accused* that the working hours are **over** 14 hours everyday during their internship.'

Normalization: 2 名實習女大生實習期間每天工時超過 14 小時 'Two female undergrad interns worked over 14 hours during their internship.'

Qualtrics Annotations: PR+:4, PS+:2

Google Annotations: PR+:5, Uu:1

If a news sentence expresses an happened event, it is expected that readers will have higher confidence in believing the event certainly happened. Adverbs such as 已 'already' or 已經 'already' which denotes an event happened in the past, are found to have an impact in increasing readers' confidence while making judgments. However, even with this kind of past-event marker, the selection of ESPs would have stronger influence to readers' commitments to the event. For instance, with the ESP 認定 'affirm' in (14), most annotators tagged the event as CT+, as compared to the ESP 指出 'point out' in (15) where annotators have less confidence in believing the event happened.

(14)
經濟部標檢局
Bureau of Standards, Metrology and Inspection
認定 這 三項 管理 系統 已
affirm the three management system already
失效
down
'The Bureau of Standards, Metrology and Inspection *affirms* that the three management systems are already **down**.'

Normalization: 這三項管理系統已失效 'The three management systems are already down.'

Qualtrics Annotations: CT+:4, PR+:2

Google Annotations: CT+:4, PS+:1, Uu:1

(15) 彭博 專欄 *指出*， 已 請
Bloomberg column point out already invite
美國國務院 **介入**。
U.S. Department of State get involved in
'The Bloomberg column *pointed out* the U.S. Department of State was already invited to **get involved in** (the event).'

Normalization: 已請美國國務院介入 'The U.S. Department of State was already invited to get involved in (the event).'

Qualtrics Annotations: CT+:1, PR+:5

Google Annotations: CT+:3, PR+:1, PS+:1, Uu:1

In addition, it is found that the source of information introduced by an ESP is less effective to Chinese readers' veridicality judgments, even if the source of information is an authority, as shown in (16). The example also presents that negation markers (e.g., 未 'no') do not always trigger readers to make negative judgments.

(16) 當地 派出所 范姓 副所長
local police station Mr. Fan Deputy Director
說，阿明 已 未 再 **接觸** 毒品
said A-Ming already no again expose drugs
'The Deputy Director of the local police station Mr. Fan *said* that A-Ming has no longer **exposed** himself to drugs.'

Normalization: 阿明還有接觸毒品 'A-Ming is still exposing himself to drugs.'

Qualtrics Annotations: CT+:1, PR+:2, PS+:1, CT−:1, PR−:1

Google Annotations: PS+:1, CT−:2, Uu:3

Theoretically, it is expected that modality markers such as 確實 'certainly', 一定 'absolutely', and 絕對 'absolutely' express high degree of certainty; and 可能 'possible' and 應該 'should' convey lower degree of confidence. However, as observed from the datasets, modality markers with high degree of certainty do not necessarily trigger readers to have the same degree of judgments. For example, despite that 確實 'certainly' is used in the two sentences listed below, the high confidence ESP 坦承 'admit' introduced in (17) would affect most annotators to fully believe the event certainly happened, as compared to the ESP 表示 'express' in (18) which annotators have less confidence. In addition, from the notion that Chinese modality markers imply a semantic source (Hsieh, 2005), if modality markers are less effective to veridicality judgments, then so are the source of information. This inference is further examined in the above findings which both source of information and modality markers do not present a clear influence to readers' veridicality judgments.

(17) 曾男 *坦承* 確實 有 **3P 性交易**
Mr. Tseng admit certainly have 3P sex trade
'Mr. Tseng *admitted* it is certain that he had a **3P sex trade**.'

Normalization: 曾男有 3P 性交易 'Mr. Tseng had a 3P sex trade.'

Qualtrics Annotations: CT+:4, PR+:2

Google Annotations: CT+:5, PR:1

(18) 憲哥 *表示*
Jacky Wu express
雙 J 戀 *確實* **存在**
Jay & Jolin love relationship certainly exist

'Jacky Wu *expressed* it is certain that the Jay & Jolin love relationship **existed**.'

Normalization: 雙 J 戀存在 'The Jay & Jolin love relationship existed.'

Qualtrics Annotations: PR+:2, PS+:4

Google Annotations: CT+:1, PR:3, PS+:1, Uu:1

Some adverbs also carry a sense of possibility, such as 有時候 'sometimes', 好像 'seem', 約 'around', 恐怕 'be afraid that', 有機會 'stand a chance' and 幾乎 'almost'. These adverbs may also show an impact on making less confident judgments as presented in (19) and (20).

(19) 民眾 報案 *指稱*， 新民路
People report a case indicate Xin-Min Road
一帶 好像 **瓦斯外洩**。
area seems gas leak
'People report a case *indicating* that there seems to be a **gas leak** on the Xin-Min Road.'

Normalization: 新民路一帶瓦斯外洩 'There seems to be a gas leak on the Xin-Min Road.'

Qualtrics Annotations: PS+:6

Google Annotations: PS+:2, Uu:4

(20) 甚至 有 法人 *預估*
even there is Juridical Person estimate
Switch 銷量 有機會 **破** 千萬
Switch sales have a chance exceed 10 million
'There is even a juridical person *estimates* that the sales of Switch has a chance of **exceeding** 10 million dollars.'

Normalization: Switch 銷量將破千萬 'The sales of Switch will exceed 10 million dollars.'

Qualtrics Annotations: PS+:6

Google Annotations: PR+:1, PS+:2, Uu:3

5. Conclusion

From the Chinese datasets collected from readers in Taiwan, it is observed that readers' veridicality judgments would be affected by whether an event is introduced by a high confidence ESP. Additionally, the source of information introduced by an ESP is less effective for readers in Taiwan to make judgments. In other words, even if the source of information is an authority, it may not increase readers' degree of confidence in believing the event happened. It is also explored that modality markers behave differently while placing into context. Modality markers with high certainty do not always trigger readers to have the same degree of confidence as proposed in linguistic theories. This finding coincides with the statement illustrated by Hsieh (2005), which shows that Chinese modality markers imply a semantic source. Other features such as temporal information, statistics, and adverbs would have an impact on readers' veridicality judgments as well. The Qualtrics dataset used in this study is released as the Chinese PragBank for further analysis.

6. Acknowledgements

The collection of the Qualtrics dataset is funded by the Ministry of Science and Technology grant No. MOST 107-2410-H-002 -162 -. Thanks for the feedback from Marie-Catherine de Marneffe while collecting the dataset, and the assistance from Chiung-Yu Chiang for distributing the compensation fee to annotators. Thanks for the reviewers' comments and suggestions to this study. Most importantly, thanks for all annotators who participated in this study.

7. Bibliographical References

Aikhenvald, A. Y. (2004). *Evidentiality*. Oxford University Press.

Banfield, A. (1982). *Unspeakable Sentences: Narration and representation in the language of fiction*. Boston : Routledge & Kegan Paul.

de Marneffe, M.-C., Manning, C. D., and Potts, C. (2012). Did it happen? The pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38(2):301–333.

Finn, R. H. (1970). A note on estimating the reliability of categorical data. *Educational and Psychological Measurement*, 30(1):70–76.

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions*. John Wiley, 2nd edition.

Giannakidou, A. (1999). Affective dependencies. *Linguistics and Philosophy*, 22(4):367–421.

Horn, L. R. (1989). *A Natural History of Negation*. University of Chicago Press.

Hsieh, C.-L. (2005). Modal verbs and modal adverbs in Chinese: An investigation into the semantic source. *UST Working Papers in Linguistics*, 1:31–58.

Huang, C.-R., Hsieh, S.-K., and Chen, K.-J. (2017). *Mandarin Chinese Words and Parts of Speech: Corpus-based Foundational Studies*. Taylor & Francis.

Hwang, . Y. (1999). Hanyu nengyuan dongci zhi yuyi yanjiu 漢語能願動詞之語義研究 'A semantic study of modal verbs in Chinese'. Master's thesis, Guoli Taiwan Shifan Daxue 國立臺灣師範大學 'National Taiwan Normal University'.

Jakobson, R. (1957). *Shifters, Verbal Categories, and the Russian Verb*. Harvard University.

Krippendorff, K. (1980). *Content Analysis: An introduction to its methodology*. Sage.

Lee, K., Artzi, Y., Choi, Y., and Zettlemoyer, L. (2015). Event detection and factuality assessment with non-expert supervision. In *Proceeding of the International Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1643–1648. Association for Computational Linguistics.

Lyons, J. (1977). *Semantics*, volume 2. Cambridge University Press.

Ma, L. (2011). Study on evidentiality in spoken Mandarin Chinese. In *6th IEEE Joint International Information Technology and Artificial Intelligence Conference*, pages 283–287.

Minard, A.-L., Marchetti, A., and Speranza, M. (2014). Event factuality in Italian: Annotation of news stories from the Ita-TimeBank. In *Proceedings of the First Italian Conference on Computational Linguistics (CLiC-it 2014)*, pages 260–264.

Mushin, I. (2001). *Evidentiality and Epistemological Stance*. John Benjamins Publishing Company.

Palmer, F. R. (2001). *Mood and Modality*. Cambridge University Press, 2 edition.

Pustejovsky, J., no, J. C., Ingria, R., Saurí, R., Gaizauskas, R., Setzer, A., and Katz, G. (2003a). Timeml: a specification language for temporal and event expressions. In *International Workshop of Computational Semantics*. Kluwer Academic Publishers.

Pustejovsky, J., no, J. C., Ingria, R., Saurí, R., Gaizauskas, R., Setzer, A., and Katz, G. (2003b). TimeML: Robust specification of event and temporal expressions in text. In *Fifth International Workshop on Computational Semantics*.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. Longman.

Robinson, W. S. (1957). The statistical measurement of agreement. *American Sociological Review*, 22(1):17–25.

Rubin, V. L., Liddy, E. D., and Kando, N., (2006). *Computing Attitude and Affect in Text: Theory and Applications*, volume 20 of *The Information Retrieval Series*, chapter Certainty Identification in Texts: Categorization Model and Manual Tagging Results, pages 61–76. Springer, Dordrecht.

Saurí, R. and Pustejovsky, J. (2009). FactBank: A corpus annotated with event factuality. *Language Resources and Evaluation*, 43(3):227–268.

Saurí, R. (2008). *A Factuality Profiler for Eventualities in Text*. Ph.D. thesis, Brandeis University.

Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428.

Steele, S., Akmajian, A., Demers, R., Jelinek, E., Kitagawa, C., Oehrle, R., and Wasow, T. (1981). *An Encyclopedia of AUX: A study in cross-linguistic equivalence*. MIT Press.

Su, Q. and Liu, P. (2012). A tentative study on the annotation of evidentiality. In *Chinese Lexical Semantics (CLSW 2012). Lecture Notes in Computer Science*, volume 7717, pages 364–372. Springer.

Su, Q., Huang, C.-R., and yun Chen, H. K. (2010). Evidentiality for text trustworthiness detection. In *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground*, pages 10–17, Uppsala, Sweden, July. Association for Computational Linguistics.

Tiee, H. H., (1985). *Studies in East Asian Linguistics*, chapter Modality in Chinese, pages 84–96. Department of East Asian Languages and Cultures, University of Southern California.

Tsang, C.-L. (1981). *A Semantic Study of Modal Auxiliary Verbs in Chinese*. Ph.D. thesis, Stanford University.

van Son, C., van Erp, M., Fokkens, A., and Vossen, P. (2014). Hope and fear: Interpreting perspectives by integrating sentiment and event factuality. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*, pages 3857–3864, May.

van Valin, R. D. and LaPolla, R. J. (1997). *Syntax: Structure, Meaning, and Function*. Cambridge University Press.

von Wright, G. H. (1951). *An Essay in Modal Logic*. North Holland.

Wiebe, J. (1994). Tracking point of view in narrative. *Computational Linguistics*, 20(2):233–287.

Zhu, . Y. (2006). Shilun xiandai hanyu de yanjuxing 試論現代漢語的言據性 'Evidential studies in modern Chinese'. *Xiandai Waiyu 現代外語 'Modern Foreign Languages'*, 4:331–337.