# How and Where do People Fail with Time: Temporal Reference Mapping Annotation by Chinese and English Bilinguals

**Yang Ye**[§], **Steven Abney**[†§]
[†]Department of Linguistics
[§]Department of Electrical Engineering and Computer Science
University of Michigan

## Abstract

This work reports on three human tense annotation experiments for Chinese verbs in Chinese-to-English translation scenarios. The results show that inter-annotator agreement increases as the context of the verb under the annotation becomes increasingly specified, i.e. as the context moves from the situation in which the target English sentence is unknown to the situation in which the target lexicon and target syntactic structure are fully specified. The annotation scheme with a fully specified syntax and lexicon in the target English sentence yields a satisfactorily high agreement rate. The annotation results were then analyzed via an ANOVA analysis, a logistic regression model and a log-linear model. The analyses reveal that while both the overt and the latent linguistic factors seem to significantly affect annotation agreement under different scenarios, the latent features are the real driving factors of tense annotation disagreement among multiple annotators. The analyses also find the verb telicity feature, aspect marker presence and syntactic embedding structure to be strongly associated with tense, suggesting their utility in the automatic tense classification task.

## 1 Introduction

In recent years, the research community has seen a fast-growing volume of work in temporal information processing. Consequently, the investigation and practice of temporal information annotation by human experts have emerged from the corpus annotation research. To evaluate automatic temporal relation classification systems, annotated corpora must be created and validated, which motivates experiments and research in temporal information annotation.

One important temporal relation distinction that human beings make is the temporal reference distinction based on relative positioning between the following three time parameters, as proposed by (Reichenbach, 1947): speech time (S), event time (E) and reference time (R). Temporal reference distinction is linguistically realized as tenses. Languages have various granularities of tense representations; some have finer-grained tenses or aspects than others. This poses a great challenge to automatic cross-lingual tense mapping. The same challenge holds for cross-lingual tense annotation, especially for language pairs that have dramatically different tense strategies. A decent solution for cross-lingual tense mapping will benefit a variety of NLP tasks such as Machine Translation, Cross-lingual Question Answering (CLQA), and Multi-lingual Information Summarization. While automatic cross-lingual tense mapping has recently started to receive research attention, such as in (Olsen,et al., 2001) and (Ye, et al., 2005), to the best of our knowledge, human performance on tense and aspect annotation for machine translation between English and Chinese has not received any systematic investigation to date. Cross-linguistic NLP tasks, especially those requiring a more accurate tense and aspect resolution, await a more focused study of human tense and aspect annotation performance.

Chinese and English are a language pair in which tense and aspect are represented at different levels of units: one being realized at the word level and the other at the morpheme level.

This paper reports on a series of cross-linguistic tense annotation experiments between Chinese and English, and provides statistical inference for different linguistic factors via a series of statistical modeling. Since tense and aspect are morphologically merged in English, tense annotation



*Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006*, pages 13–20,
Sydney, July 2006. 

discussed in this paper also includes elements of aspect. We only deal with tense annotation in Chinese-to-English scenario in the scope of this paper.

The remaining part of the paper is organized as follows: Section 2 summarizes the significant related works in temporal information annotation and points out how this study relates to yet differs from them. Section 3 reports the details of three tense annotation experiments under three scenarios. Section 4 discusses the inter-judge agreement by presenting two measures of agreement: the Kappa Statistic and accuracy-based measurement. Section 5 investigates and reports on the significance of different linguistic factors in tense annotation via an ANOVA analysis, a logistic regression analysis and a log-linear model analysis. Finally, section 6 concludes the paper and points out directions for future research.

## 2 Related Work

There are two basic types of temporal location relationships. The first one is the ternary classification of past, present and future. The second one is the binary classification of "BEFORE" versus "AFTER". These two types of temporal relationships are intrinsically related but each stands as a separate issue and is dealt with in different works. While the "BEFORE" versus "AFTER" relationship can easily be transferred across a language pair, the ternary tense taxonomy is often very hard to transfer from one language to another.

(Wilson, et al., 1997) describes a multilingual approach to annotating temporal information, which involves flagging a temporal expression in the document and identifying the time value that the expression designates. Their work reports an inter-annotator reliability F-measure of 0.79 and 0.86 respectively for English corpora.

(Katz, et al., 2001) describes a simple and general technique for the annotation of temporal relation information based on binary interval relation types: precedence and inclusion. Their annotation scheme could benefit a range of NLP applications and is easy to carry out.

(Pustejovsky et al., 2004) reports an annotation scheme, the TimeML metadata, for the markup of events and their anchoring in documents. The annotation schema of TimeML is very fine-grained with a wide coverage of different event types, dependencies between events and times, as well as "LINK" tags which encode the various relations existing between the temporal elements of a document. The challenge of human labeling of links among eventualities was discussed at great length in their paper. Automatic "time-stamping" was attempted on a small sample of text in an earlier work of (Mani, 2003). The result was not particularly promising. It showed the need for a larger quantity of training data as well as more predictive features, especially on the discourse level. At the word level, the semantic representation of tenses could be approached in various ways depending on different applications. So far, their work has gone the furthest towards establishing a broad and open standard metadata mark-up language for natural language texts.

(Setzer, et al., 2004) presents a method of evaluating temporal order relation annotations and an approach to facilitate the creation of a gold standard by introducing the notion of temporal closure, which can be deduced from any annotations through using a set of inference rules.

From the above works, it can be seen that the effort in temporal information annotation has thus far been dominated by annotating temporal relations that hold entities such as events or times explicitly mentioned in the text. Cross-linguistic tense and aspect annotation has so far gone unstudied.

## 3 Chinese Tense Annotation Experiments[1]

In current section, we present three tense annotation experiments with the following scenarios:

1. Null-control situation by native Chinese speakers where the annotators were provided with the source Chinese sentences but not the English translations;

2. High-control situation by native English speakers where the annotators were provided with the Chinese sentences as well as English translations with specified syntax and lexicons;

3. Semi-control situation by native English speakers where the annotators were allowed to choose the syntax and lexicons for the English sentence with appropriate tenses;

[1] All experiments in the paper are approved by Behavioral Sciences Institutional Review Board at the University of Michigan, the IRB file number is B04-00007481-I.

### 3.1 Experiment One

Experiment One presents the first scenario of tense annotation for Chinese verbs in Chinese-to-English cross-lingual situation. In the first scenario, the annotation experiment was carried out on 25 news articles from LDC Xinhua News release with category number LDC2001T11. The articles were divided into 5 groups with 5 articles in each group. There are a total number of 985 verbs. For each group, three native Chinese speakers who were bilingual in Chinese and English annotated the tense of the verbs in the articles independently. Prior to annotating the data, the annotators underwent brief training during which they were asked to read an example of a Chinese sentence for each tense and make sure they understand the examples. During the annotation, the annotators were asked to read the whole articles first and then select a tense tag based on the context of each verb. The tense taxonomy provided to the annotators include the twelve tenses that are different combinations of the simple tenses (present, past and future), the prograssive aspect and the perfect aspect. In cases where the judges were unable to decide the tense of a verb, they were instructed to tag it as "unknown". In this experiment, the annotators were asked to tag the tense for all Chinese words that were tagged as verbs in the Penn Treebank corpora. Conceivably, the task under the current scenario is meta-linguistic in nature for the reason that tense is an elusive notion for Chinese speakers. Nevertheless, the experiment provides a baseline situation for human tense annotation agreement. The following is an example of the annotation where the annotators were to choose an appropriate tense tag from the provided tense tags:

((IP (NP-TPC (NP-PN (NR 中国))(NP (NN 建筑)(NN 市场)))(LCP-TMP (NP (NT 近年))(LC 来)) (NP-SBJ (NP (PP (P 对)(NP (NN 外)))(NP (NN 开放)))(NP (NN 步伐)))(VP (ADVP (AD 进一步)) (VP (***VV 加快***)))(PU 。)) )
1. simple present tense
2. simple past tense
3. simple future tense
4. present perfect tense
5. past perfect tense
6. future perfect tense
7. present progressive tense
8. past progressive tense
9. future progressive
10. present perfect progressive
11. past perfect progressive

### 3.2 Experiment Two

Experiment Two was carried out using 25 news articles from the parallel Chinese and English news articles available from LDC Multiple Translation Chinese corpora (MTC catalog number LDC2002T01). In the previous experiment, the annotators tagged all verbs. In the current experimental set-up, we preprocessed the materials and removed those verbs that lose their verbal status in translation from Chinese to English due to nominalization. After this preprocessing, there was a total of 288 verbs annotated by the annotators. Three native speakers, who were bilingually fluent in English and Chinese, were recruited to annotate the tense for the English verbs that were translated from Chinese. As in the previous scenario, the annotators were encouraged to pay attention to the context of the target verb when tagging its tense. The annotators were provided with the full taxonomy illustrated by examples of English verbs and they worked independently. The following is an example of the annotation where the annotators were to choose an appropriate tense tag from the provided tense tags:

据统计，这些城市去年***完成***国内生产总值一百九十多亿元，比开放前的一九九一年增长九成多。
According to statistics, the cities (***achieve***) a combined gross domestic product of RMB19 billion last year, an increase of more than 90% over 1991 before their opening.
A. achieves
B. achieved
C. will achieve
D. are achieving
E. were achieving
F. will be achieving
G. have achieved
H. had achieved
I. will have achieved
J. have been achieving
K. had been achieving
L. will have been achieving
M. would achieve

### 3.3 Experiment Three

Experiment Three was an experiment simulated on 52 Xinhua news articles from the Multiple Translation Corpus (MTC) mentioned in the previous section. Since in the MTC corpora, each Chinese article is translated into English by ten human translation teams, conceptually, we could view these ten translation teams as different annotators. They were making decisions about appropriate tense for the English verbs. These annotators differ from those in Experiment Two described above in that they were allowed to choose any syntactic structure and verb lexicon. This is because they were performing tense annotation in a bigger task of sentence translation. Therefore, their tense annotations were performed with much less specification of the annotation context. We manually aligned the Chinese verbs with the English verbs for the 10 translation teams from the MTC corpora and thus obtained our third source of tense annotation results. For the Chinese verbs

that were not translated as verbs into English, we assigned a "Not Available" tag. There are 1505 verbs in total including the ones that lost their verbal status across the language.

## 4 Inter-Judge Agreement

Researchers use consistency checking to validate human annotation experiments. There are various ways of performing consistency checking described in the literature, depending on the scale of the measurements. Each has its advantages and disadvantages. Since our tense taxonomy is nominal without any ordinal information, Kappa statistics measurement is the most appropriate choice to measure inter-judge agreement.

### 4.1 Kappa Statistic

Kappa scores were calculated for the three human judges' annotation results. The Kappa score is the de facto standard for evaluating inter-judge agreement on tagging tasks. It reports the agreement rate among multiple annotators while correcting for the agreement brought about by pure chance. It is defined by the following formula, where P(A) is the observed agreement among the judges and P(E) is the expected agreement:

$$k = \frac{P(A) - P(E)}{1 - P(E)} \quad (1)$$

Depending on how one identifies the expected agreement brought about by pure chance, there are two ways to calculate the Kappa score. One is the "Seigel-Castellian" Kappa discussed in (Eugenio, 2004), which assumes that there is one hypothetical distribution of labels for all judges. In contrast, the "Cohen" Kappa discussed in (Cohen, 1960), assumes that each annotator has an individual distribution of labels. This discrepancy slightly affects the calculation of P(E). There is no consensus regarding which Kappa is the "right" one and researchers use both. In our experiments, we use the "Seigel-Castellian" Kappa.

The Kappa statistic for the annotation results of Experiment One are 0.277 on the full taxonomy and 0.37 if we collapse the tenses into three big classes: present, past and future. The observed agreement rate,that is, P(A), is 0.42.

The Kappa score for tense resolution from the ten human translation teams for the 52 Xinhua news articles is 0.585 on the full taxonomy; we expect the Kappa score to be higher if we exclude the verbs that are nominalized. Interestingly, the Kappa score calculated by collapsing the 13 tenses into 3 tenses (present, past and future) is only slightly higher: 0.595. The observed agreement rate is 0.72.

Human tense annotation in the Chinese-to-English restricted translation scenario achieved a Kappa score of 0.723 on the full taxonomy with an observed agreement of 0.798. If we collapse simple past and present perfect, the Kappa score goes up to 0.792 with an observed agreement of 0.893. The Kappa score is 0.81 on the reduced taxonomy.

### 4.2 Accuracy

The Kappa score is a relatively conservative measurement of the inter-judge agreement rate. Conceptually, we could also obtain an alternative measurement of reliability by taking one annotator as the gold standard at one time and averaging over the accuracies of the different annotators across different gold standards. While it is true that numerically, this would yield a higher score than the Kappa score and seems to be inflating the agreement rate, we argue that the difference between the Kappa score and the accuracy-based measurement is not limited to one being more aggressive than the other. The policies of these two measurements are different. The Kappa score is concerned purely with agreement without any consideration of truthfulness or falsehood, while the procedure we described above gives equal weights to each annotator being the gold standard. Therefore, it considers both the agreement and the truthfulness of the annotation. Additionally, the accuracy-based measurement is the same measurement that is typically used to evaluate machine performance; therefore it gives a genuine ceiling for machine performance.

The accuracy under such a scheme for the three annotators in Experiment One is 43% on the full tense taxonomy.

The accuracy under such a scheme for tense generation agreement from three annotators in Experiment Two is 80% on the full tense taxonomy.

The accuracy under such a scheme for the ten translation teams in Experiment Three is 70.8% on the full tense taxonomy.

Table 1 summarizes the inter-judge agreement for the three experiments.

Examining the annotation results, we identified the following sources of disagreement. While the

| Agreement | Exp 1 | Exp 2 | Exp 3 |
|---|---|---|---|
| Kappa Statistic | 0.277 | 0.723 | 0.585 |
| Kappa Statistic (Reduced Taxonomy) | 0.37 | 0.81 | 0.595 |
| Accuracy | 43% | 80% | 70.8% |

Table 1: Inter-Annotator Agreement for the Three Tense Annotation Experiments

first two factors can be controlled for by a clearly pre-defined annotation guideline, the last two factors are intrinsically rooted in natural languages and therefore hard to deal with:

1. Different compliance with Sequence of Tense (SOT) principle among annotators;
2. "Headline Effect";
3. Ambiguous POS of the "verb": sometimes it is not clear whether a verb is adjective or past participle. *e.g. The Fenglingdu Economic Development Zone is the only one in China that **is/was built** on the basis of a small town.*
4. Ambiguous aspectual property of the verb: the annotator's view with respect to whether or not the verb is an atelic verb or a telic verb. *e.g. "statistics **showed/show**....."*

Put abstractly, ambiguity is an intrinsic property of natural languages. A taxonomy allows us to investigate the research problem, yet any clearly defined discrete taxonomy will inevitably fail on boundary cases between different classes.

## 5 Significance of Linguistic Factors in Annotation

In the NLP community, researchers carry out annotation experiments mainly to acquire a gold standard data set for evaluation. Little effort has been made beyond the scope of agreement rate calculations. We propose that not only does feature analysis for annotation experiments fall under the concern of psycholinguists, it also merits investigation within the enterprise of natural language processing. There are at least two ways that the analysis of annotation results can help the NLP task besides just providing a gold standard: identifying certain features that are responsible for the inter-judge disagreement and modeling the situation of associations among the different features. The former attempts to answer the

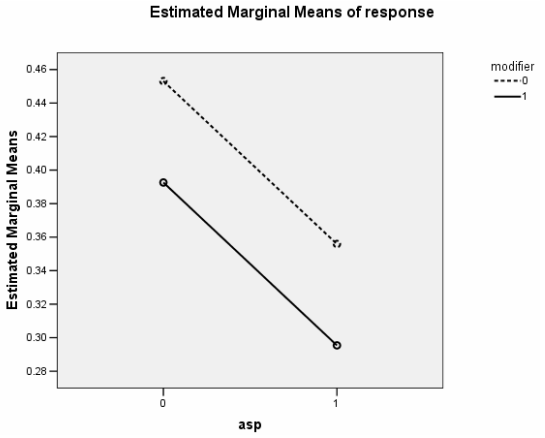


Figure 1: Interaction between Aspect Marker and Temporal Modifier

question of where the challenge for human classification comes from, and thereby provides an external reference for an automatic NLP system, although not necessarily in a direct way. The latter sheds light on the structures hidden among groups of features, the identification of which could provide insights for feature selection as well as offer convergent evidence for the significance of certain features confirmed from classification practice based on machine learning.

In this section, we discuss at some length a feature analysis for the results of each of the annotation experiments discussed in the previous sections and summarize the findings.

### 5.1 ANOVA analysis of Agreement and Linguistic Factors in Free Translation Tense Annotation

This analysis tries to find the relationship between the linguistic properties of the verb and the tense annotation agreement across the ten different translation teams in Experiment Three. Specifically, we use an ANOVA analysis to explore how the overall variance in the inconsistency of the tenses of a particular verb with respect to different translation teams can be attributed to different linguistic properties associated with the Chinese verb. It is a three-way ANOVA with three linguistic factors under investigation: whether the sentence contains a temporal modifier or not; whether the verb is embedded in a relative clause, a sentential complement, an appositive clause or none of the above; and whether the verb is followed by aspect markers or not. The dependent variable is the inconsistency of the tenses from the teams. The

inconsistency rate is measured by the ratio of the number of distinct tenses over the number of tense tokens from the ten translation teams.

Our ANOVA analysis shows that all of the three main effects, i.e. the embedding structures of the verb ($p \ll 0.001$), the presence of aspect markers ($p \ll 0.01$), and the presence of temporal modifiers ($p < 0.05$) significantly affect the rate of disagreement in tense generation among the different translation teams. The following graphs show the trend: tense generation disagreement rates are consistently lower when the Chinese aspect marker is present, whether there is a temporal modifier present or not (Figure 1). The model also suggested that the presence of temporal modifiers is associated with a lower rate of disagreement for three embedding structures except for verbs in sentential complements (Figure 2, 0: the verb is not in any embedding structures; 1: the verb is embedded in a relative clause; 2: the verb is embedded in an appositive clause; 3: the verb is embedded in sentential complement). Our explanation for this is that the annotators receive varying degrees of prescriptive writing training, so when there is a temporal modifier in the sentence as a confounder, there will be a larger number, a higher incidence of SOT violations than when there is no temporal modifier present in the sentence. On top of this, the rate of disagreement in tense tagging between the case where a temporal modifier is present in the sentence and the case where it is not depends on different types of embedding structures (Figure 2, $p$ value $< 0.05$).

We also note that the relative clause embedding structure is associated with a much higher disagreement rate than any other embedding structures (Figure 3).

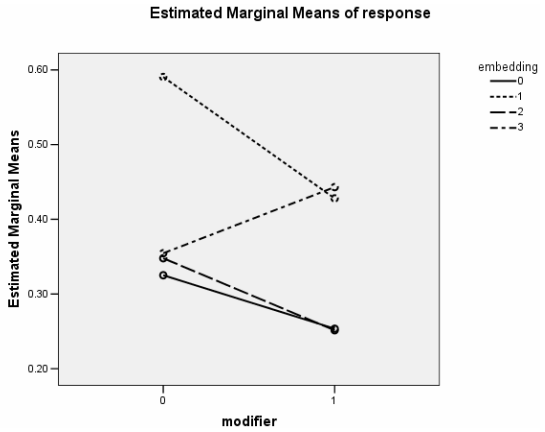


Figure 2: Interaction between the Temporal Modifier and the Syntactic Embedding Structure

### 5.2 Logistic Regression Analysis of Agreement and Linguistic Factors in Restricted Tense Annotation

The ANOVA analysis in the previous section is concerned with the confounding power of the overt linguistic features. The current section examines the significance of the more latent features on tense annotation agreement when the SOT effect is removed by providing the annotators a clear guideline about the SOT principle. Specifically, we are interested in the effect of verb telicity and punctuality features on tense annotation agreement. The telicity and punctuality features were obtained through manual annotation based on the situation in the context. The data are from Experiment Two. Since there are only three annotators, the inconsistency rate we discussed in 5.1 would have insufficient variance in the current scenario, making logistic regression a more appropriate analysis. The response is now binary being either agreement or disagreement (including partial agreement and pure disagreement). To avoid a multi-colinearity problem, we model Chinese features and English features separately. In order to truly investigate the effects of the latent features, we keep the overt linguistic features in the model as well. The overt features include: type of syntactic embedding, presence of aspect marker, presence of temporal expression in the sentence, whether the verb is in a headline or not, and the presence of certain signal adverbs including "yijing"(already), "zhengzai" (Chinese pre-verb progressive marker), "jiang"(Chinese pre-verbal adverb indicating future tense). We used backward elimination to obtain the final model.

The result showed that punctuality is the only factor that significantly affects the agreement rate among multiple judges in both the model of English features and the model of Chinese features. The significance level is higher for the punctuality of English verbs, suggesting that the source language environment is more relevant in tense generation. The annotators are roughly four times more likely to fail to agree on the tense for verbs associated with an interval event. This supports the hypothesis that human beings use the latent features for tense classification tasks. Surprisingly, the telicity feature is not significant at all. We sus-

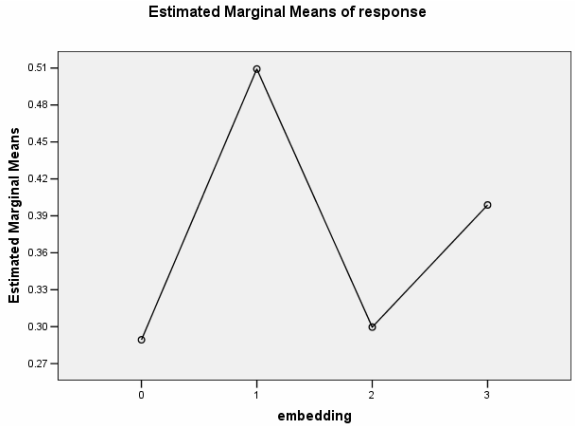


Figure 3: Effect of Syntactic Embedding Structure on Tense Annotation Disagreement

pect this is partly due to the correlation between the punctuality feature and the telicity feature. Additionally, none of the overt linguistic features is significant in the presence of the latent features, which implies that the latent features drive disagreement among multiple annotators.

### 5.3 Log-linear Model Analysis of Associations between Linguistic Factors in Free Translation Tense Annotation

This section discusses the association patterns between tense and the relevant linguistic factors via a log-linear model. A log-linear model is a special case of generalized linear models (GLMs) and has been widely applied in many fields of social science research for multivariate analysis of categorical data. The model reveals the interaction between categorical variables. The log-linear model is different from other GLMs in that it does not distinguish between "response" and "explanatory variables". All variables are treated alike as "response variables", whose mutual associations are explored. Under the log-linear model, the expected cell frequencies are functions of all variables in the model. The most parsimonious model that produces the smallest discrepancy between the expected cell and the observed cell frequencies is chosen as the final model. This provides the best explanation of the observed relationships among variables.

We use the data from Experiment Two for the current analysis. The results show that three linguistic features under investigation are significantly associated with tense. First, there is a strong association between aspect marker presence and tense, independent of punctuality, telicity feature and embedding structure. Second, there is a strong association between telicity and tense, independent of punctuality, aspect marker presence and punctuality feature. Thirdly, there is a strong association between embedding structure and tense, independent of telicity, punctuality feature and aspect marker presence. This result is consistent with (Olsen, 2001), in that the lexical telicity feature, when used heuristically as the single knowledge source, can achieve a good prediction of verb tense in Chinese to English Machine Translation. For example, the odds of the verb being atelic in the past tense is 2.5 times the odds of the verb being atelic in the future tense, with a 95% confidence interval of (0.9, 7.2). And the odds of a verb in the future tense having an aspect marker approaches zero when compared to the odds of a verb in the past tense having an aspect marker.

Putting together the pieces from the logistic analysis and the current analysis, we see that annotators fail to agree on tense selection mostly with apunctual verbs, while the agreed-upon tense is jointly decided by the telicity feature, aspect marker feature and the syntactic embedding structure that are associated with the verb.

## 6 Conclusions and Future Work

As the initial attempt to assess human beings' cross-lingual tense annotation, the current paper carries out a series of tense annotation experiments between Chinese and English under different scenarios. We show that even if tense is an abstract grammatical category, multiple annotators are still able to achieve a good agreement rate when the target English context is fully specified. We also show that in a non-restricted scenario, the overt linguistic features (aspect markers, embedding structures and temporal modifiers), can cause people to fail to agree with each other significantly in tense annotation. These factors exhibit certain interaction patterns in the decision making of the annotators. Our analysis of the annotation results from the scenario with a fully specified context show that people tend to fail to agree with each other on tense for verbs associated with interval events. The disagreement seems not to be driven by the overt linguistic features such as embedding structure and aspect markers. Lastly, among a set of overt and latent linguistic features, aspect marker presence, embedding structure and

the telicity feature exhibit the strongest association with tense, potentially indicating their high utility in tense classification task.

The current analysis, while suggesting certain interesting patterns in tense annotation, could be more significant if the findings could be replicated by experiments of different scales on different data sets. Furthermore, the statistical analysis could be more finely geared to capture the more subtle distinctions encoded in the features.

**Acknowledgement** All of the annotation experiments in this paper are funded by Rackham Graduate School's Discretionary Funds at the University of Michigan.

## References

Hans Reichenbach,1947. Elements of Symbolic Logic, Macmillan, New York, N.Y.

Mari Olson, David Traum, Carol Van Ess-Dykema, and Amy Weinberg, 2001. Implicit Cues for Explicit Generation: Using Telicity as a Cue for Tense Structure in a Chinese to English MT System, Proceedings Machine Translation Summit VIII, Santiago de Compostela, Spain.

Yang Ye, Zhu Zhang, 2005. Tense Tagging for Verbs in Cross-Lingual Context: A Case Study. Proceedings of 2nd International Joint Conference in Natural Language Processing (IJCNLP), 885-895.

George Wilson, Inderjeet Mani, Beth Sundheim, and Lisa Ferro, 2001. A Multilingual Approach to Annotating and Extracting Temporal Information, Proceedings of the ACL 2001 Workshop on Temporal And Spatial Information Processing, 39th Annual Meeting of ACL, Toulouse, 81-87.

Graham Katz and Fabrizio Arosio, 2001. The Annotation of Temporal Information in Natural Language Sentences, Proceedings of the ACL 2001 Workshop on Temporal And Spatial Information Processing, 39th Annual Meeting of ACL, Toulouse, 104-111.

James Pustejovsky, Robert Ingria, Roser Sauri, Jose Castano, Jessica Littman, Rob Gaizauskas, Andrea Setzer, Graham Katz, and Inderjeet Mani. 2004. The Specification Language TimeML. The Language of Time: A Reader. Oxford, 185-96.

Barbara Di Eugenio and Michael Glass. 2004. The kappa statistic: A second look. Computational Linguistics, 30(1): 95-101.

Inderjeet Mani, 2003. Recent Developments in Temporal Information Extraction. In Nicolov, N. and Mitkov, R., editors, Proceedings of RANLP'03. John Benjamins.

Andrea Setzer, Robert Gaizauskas, and Mark Hepple, 2003. Using Semantic Inferences for Temporal Annotation Comparison, Proceedings of the Fourth International Workshop on Inference in Computational Semantics (ICOS-4), INRIA, Lorraine, Nancy, France, September 25-26, 185-96.

Jacob Cohen, 1960. A Coefficient of Agreement for Nominal Scales, Educational and Psychological Measurement, 20, 37-46.