UG-schematic Annotation for Event Nominals: A Case Study in Mandarin Chinese

Wenxi Li
Peking University, Department of Chinese Language and Literature
liwenxi@pku.edu.cn

Yutong Zhang
Peking University, Department of Chinese Language and Literature
1900014129@pku.edu.cn

Guy Emerson
University of Cambridge, Department of Computer Science and Technology
gete2@cam.ac.uk

Weiwei Sun
University of Cambridge, Department of Computer Science and Technology
ws390@cam.ac.uk

Divergence of languages observed at the surface level is a major challenge encountered by multilingual data representation, especially when typologically distant languages are involved. Drawing inspiration from a formalist Chomskyan perspective towards language universals, Universal Grammar (UG), this article uses deductively pre-defined universals to analyze a multilingually heterogeneous phenomenon, event nominals. In this way, deeper universality of event nominals beneath their huge divergence in different languages is uncovered, which empowers us to break barriers between languages and thus extend insights from some synthetic languages to a non-inflectional language, Mandarin Chinese. Our empirical investigation also demonstrates this UG-inspired schema is effective: With its assistance, the inter-annotator agreement (IAA) for identifying event nominals in Mandarin grows from 88.02% to 94.99%, and automatic detection of event-reading nominalizations on the newly-established data achieves an accuracy of 94.76% and an F_1 score of 91.3%, which significantly surpass those achieved on the pre-existing resource by 9.8% and 5.2%, respectively. Our systematic analysis also sheds light on nominal semantic

Action Editor: Xuanjing Huang. Submission received: 1 August 2023; revised version received: 11 November 2023; accepted for publication: 29 November 2023.

https://doi.org/10.1162/coli_a_00504

role labeling. By providing a clear definition and classification on arguments of event nominal, the IAA of this task significantly increases from 90.46% to 98.04%.

1. Introduction

Universal Dependencies (UD; Nivre et al. 2016, 2020) is a linguistic framework and annotation scheme that aims to provide a consistent and standardized representation for morpho-syntactic information of different languages. Aligning itself with the claim of Lexical Functional Grammar (LFG; Bresnan 2001) that universality can be established on grammatical functions, UD centrally captures relations between two fundamental linguistic units, nominals (canonically used to represent entities) and clauses (canonically used to represent events). In this manner, it provides a framework for a basic worldview namely, how entities participate in events, and thus is thought to have the capability to perform annotations on multilingual data.

However, when we extend the scope of events beyond the core of clauses to encompass event nominals, it is difficult to maintain a UD-style analysis. The challenge firstly comes from diverse morpho-syntactic characteristics exhibited by event nominalizations across languages. As shown by Figure 1, synthetic languages like English typically use various functional morphemes to express different degrees of nominalizations, for example, the derivative noun *publication*, the gerund *studying*, and the clausal complement *that she is talented in studying linguistics*. Isolating languages like Mandarin, on the other hand, rarely rely on such grammatical markers. Moreover, the relational architecture of UD may also be challenged by multilingual heterogeneity of argumenthood associated with event nominals. Different languages may realize semantically similar participants of event nominals, such as *this paper*, *linguistics*, and their Mandarin counterparts 这篇文章, 语言学 through various syntactic constructions, which complicates the nominal Semantic Role Labeling (SRL) task.

Table 1 further provides a systematic overview of morpho-syntactic features associated with event nominals and their arguments. It demonstrates that there exist considerable variations in typologically distant languages with regard to whether these features are realized as pronounced morphemes. We argue that such multilingual heterogeneity of event nominals is not an isolated example but part of a common problem of how

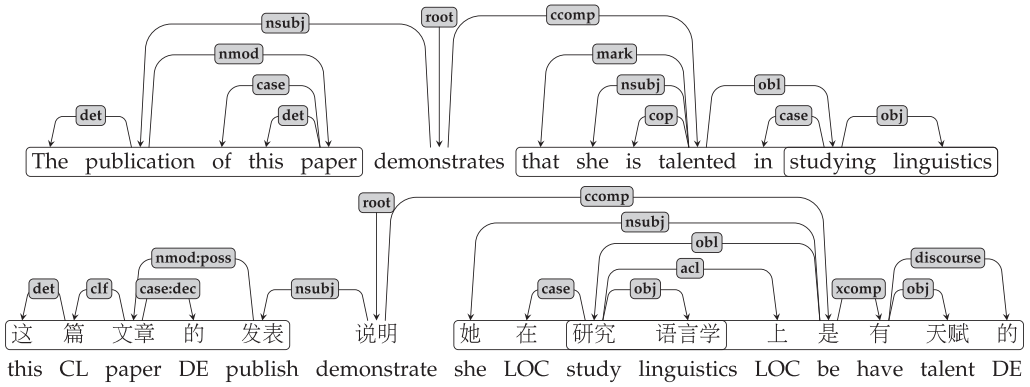

Figure 1
UD-style analysis on a English-Chinese sentence pair. Event-nominals-related phrases are circled.

Table 1
Verb- and noun-related features of event nominals, adapted from Alexiadou et al. (2011). '+' means the language has this feature and '−' means not. In English, nominative case, accusative case, and gender only apply to personal pronouns so they are marked by '?'.

	Mandarin	English	German
Subject with nominative case	−	?	+
Accusative case on object	−	?	+
Projection of outer Aspect	−	+	+
Modal or auxiliary verb	+	+	+
Complementizer	−	+	+
Verb suffix	−	+	+
Genitive/PP-subject	−	+	+
Genitive/PP-object	−	+	+
Gender	−	?	+
Quantity	−	+	+
Determiner	−	+	+
Noun suffix	−	+	+

we can accommodate multilingual phenomena with significant morpho-syntactic diversity in a unified way, or more theoretically speaking, how we can uncover intrinsic commonalities of world's languages beneath their surface-level variations. We delve into explorations of theoretical linguists in language universals to derive guidance and inspiration.

There are two main, seemingly opposing, perspectives towards universals in theoretical linguistics. The first perspective is rooted in formalist Chomskyan theories and asserts that universals are an inherent aspect of the human cognitive system, present in all individuals from birth. This innate faculty, termed as Universal Grammar (UG), is believed to encompass a set of pre-programmed rules that underlie the structural properties of languages (Chomsky 1965, 2007). The second perspective aligns with functionalist typology and suggests that universals manifest as systematic patterns observed across different languages (Greenberg 1963). These two perspectives also give rise to distinct approaches in the discovery of universals. The universal within the formalist perspective necessitates the adoption of a certain UG theory to deduce the pre-existing rules. In contrast, the functionalist-typological one requires inductive identification based on extensive empirical evidence.

To the best of our knowledge, existing research in NLP only adheres to the functionalist typological universal within the inductive paradigm. Utilizing either vectors (Huang et al. 2019; Devlin et al. 2019; Conneau et al. 2020) or discrete symbols (Abend and Rappoport 2013; Abzianidze and Bos 2017; Nivre et al. 2016, 2020), these studies strive to establish unified representations for substantial real-world texts in multiple languages. Nonetheless, as suggested by challenges mentioned before, inductively describing observable universals inevitably encounters obstacles presented by morpho-syntactic heterogeneity at the surface level.

This work alternatively explores the possibility of incorporating the formalist universal as well as the deductive paradigm in the field of NLP (§2). Specifically, our examination is tied to a particular UG model, namely, the Minimalist Approach (Chomsky 1993, 1995). This post-1990s generative approach recognizes functional rather than content words as syntactic heads and posits that their cognitive-semantic functions

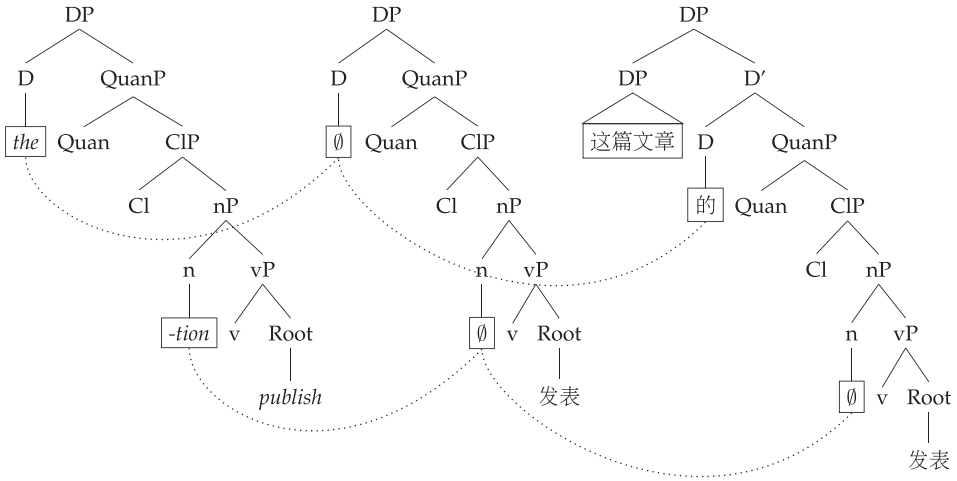

Figure 2
Analysis for *the publication*, 发表(*the publication*), and 这篇文章的发表(*the publication of this paper*). "∅" indicates an unpronounced grammatical marker.

are predefined language universals, irrespective of diverse surface forms (Wiltschko 2014; Ramchand and Svenonius 2014). With regard to event nominals, this theoretical assumption could cover morpho-syntactic features in Table 1 through multilingually compatible functions like categorizing, viewing, anchoring, and linking events or entities, which enables a unified explanation of event nominals as the interruption of event-related functional projections by entity-related ones regardless of language typology. Consequently, as exemplified by Figure 2, analysis on event nominals in genetically related Indo-European languages where there are rich functional morphemes can be extended to non-inflectional languages as they are linked by the universal functions—we can assume the existence of functional projections in Mandarin event nominals though they are not triggered by pronounced markers like *the* and *-tion* in English. What is more, *this paper*, which is realized as a PP-object, could be used as a reference for analyzing syntactic status of its counterpart 这篇文章. In this sense, within this UG-inspired framework, we could even benefit from high heterogeneity of multilingual event nominals rather than being hampered by it.

We further present the feasibility and validity of the UG-schematic representation in NLP research by applying it to event nominals in Mandarin. After introducing the schema, nominalizations in Mandarin can be clearly defined and classified, which firstly contributes to the identification task. Compared with the initial inter-annotator agreement (IAA) of 88.02%, manual annotation assisted by the UG-inspired schema achieves a relatively higher IAA of 94.99%. Furthermore, the automated tagging process demonstrates commendable performance on the newly annotated data, attaining an accuracy of 98.79% and an F_1 score of 99.71%. Specific to event-reading nominalizations, an accuracy of 94.76% and a F_1 score of 91.3% are obtained, which are remarkably higher than those on an existing resource by 9.8% and 5.2%, respectively (see §3). Moreover, syntactic analysis underlying the classification system provides a precise definition of arguments and categorizes them into either syntactically fixed or non-syntactically fixed ones. As corroborated by subsequent annotation experiments in nominal SRL (§4), this distinction not only signifies varying levels of annotation complexity, but also offers

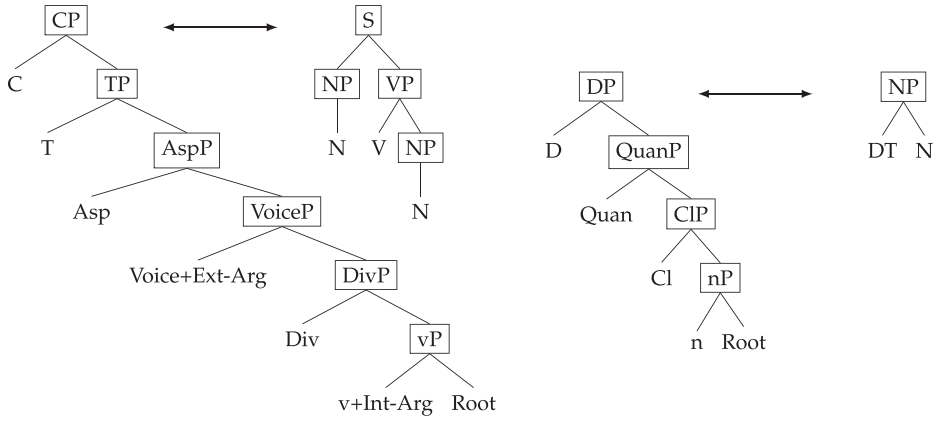

Figure 3
Syntactic schema of events/clauses and entities/noun phrases within the Minimalist theories, with a comparison to traditional surface-oriented phrase-structure analysis. Maximal projections of heads that are circled include: the complementizer phrase CP, the tense phrase TP, the aspect phrase AspP, the voice phrase VoiceP, the division phrase DivP, the verb-categorizer phrase vP, the sentence S, the noun phrase NP, the verb phrase VP, the determiner phrase DP, the quantity phrase QuanP, the classifier phrase ClP, and the noun-categorizer phrase nP. Ext-Arg means the external argument and Int-Arg means the internal argument. It can be seen that in surface-oriented trees other than DT, the determiner, which is always expressed by a free morpheme and used as a decoration of the noun, all the other bounded functional morphemes, which are heads in the Minimalist Approach, are not examined or visualized.

valuable insights into potential strategies for mitigating these difficulties. The quality of annotation on non-adverbial arguments of event nominals could be improved from 90.46% to 98.04% by involving more syntactic and contextual information.

We believe all the achievements above indicate shifts in this work, which are from predicate–argument structures to event structures, from a bottom–up description to a top–down predefinition, as well as from the grammatical functions of observable content words to those of non-compulsory functional morphemes, offers a meaningful and practical solution for uncovering language universals.

2. A UG-inspired Schema on Event Nominals of Mandarin Chinese

As shown by Figure 3, the Minimalisit Approach treats functional items as syntactic heads and content words are regarded as their decorations. Also, each of them has a unique cognitive-semantic function, including categorizing, viewing, anchoring, and linking to discourses (see Table 2), which is cross-lingually universal (Wiltschko 2014; Ramchand and Svenonius 2014).

Table 2
Functional heads and their cognitive-semantic functions of events and entities, based on Wiltschko (2014).

	Discourse link	Anchor	View	Categorize
event	CP	TP/AspP	DivP	vP
entity	DP	QuanP	ClP	nP

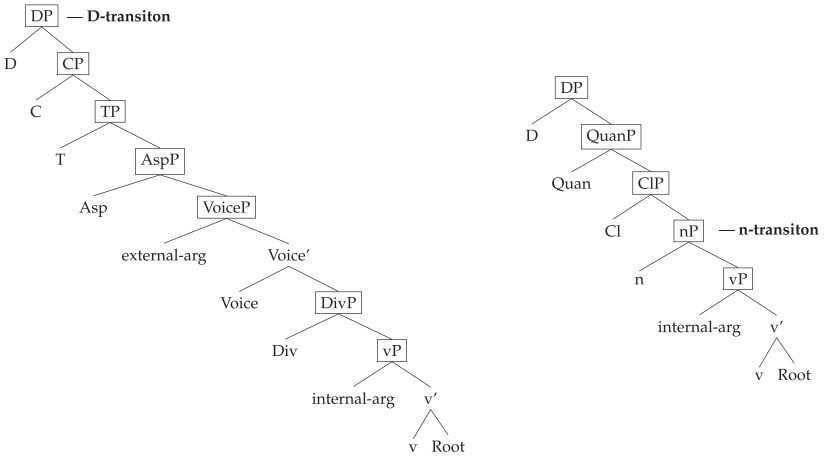

Figure 4
Syntactic schema for D-transition/n-transition adapted from Alexiadou (2020). D-transited event nominals involve more event-related projections and perform more like verbs, but are finally governed by the determiner phrase and used as complements, while the n-transition-marker forbids event-related projections earlier and thus makes event nominals more similar to nouns.

Drawing inspiration from the theoretical assumption that universals can be established on the functions of functional categories, we believe that it is plausible to extend insights provided by inflection-style markers to a typologically-distant language where there are almost no explicit markers. In particular, we consider Alexiadou's (2020) system, which is evidenced by rich functional morphemes in Indo-European languages and then divides event nominals into two types, *D-transition* and *n-transition*, as a UG-inspired schema. With a slight modification, the schema (as illustrated in Figure 4) is applicable to Mandarin as exemplified by the following examples and their English counterparts.

D-transition: Fact Reading. The D-transition pertains to actualized events, commonly referred to as facts. We thus designate this type of event nominal as having a *fact* reading. Despite functioning as clausal complements, this type of event nominal in Mandarin could still be tagged as VV, akin to typical verbs, because both of them allow for relatively complete event-related projections (see Figure 5). From a formal semantic perspective, the event variable denoted by this kind of nominal event is bounded and applicable only within a specific context.

n-transition: Event Reading. The n-transition acts on unactualized, unbounded, and general events and thus we refer to it as having an *event* reading. Depending on the presence of internal arguments, event-reading nominal events in Mandarin can be further classified into two subtypes. The former case is constrained by the internal argument acting as a modifier whereas the latter solely denotes the event itself (see Figure 6). Both subtypes are tagged as VN.

Entity Reading. It is noteworthy that certain events in Mandarin have undergone complete semantic transformations to denote entities. They are always the outcome or result

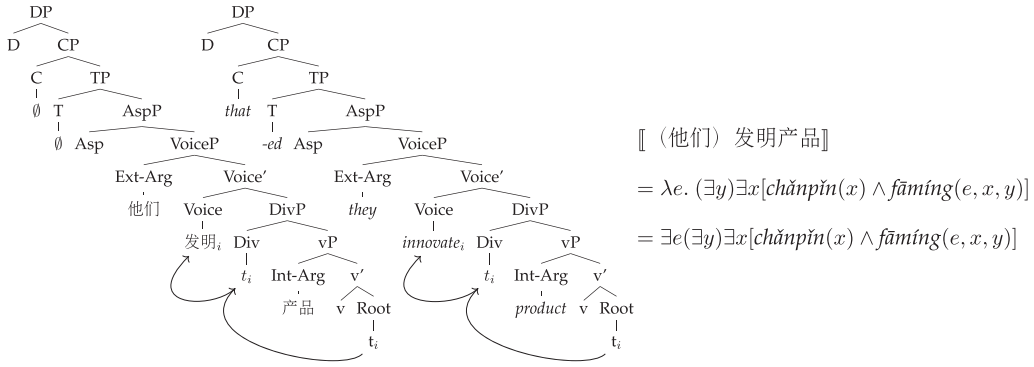

Figure 5
Syntactic and semantic analysis towards fact-reading event nominals. 他们(tāmen) means *they*; 发明(fāmíng) means *innovate*, and 产品(chǎnpǐn) means *products*.

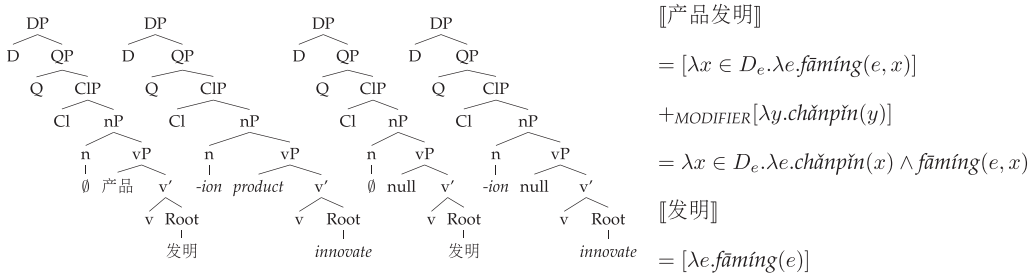

Figure 6
Syntactic and semantic analysis towards the two subtypes of event-reading event nominals. The first subtype has an internal argument as a modifier and can be translated to *product innovation* or *innovation of product*. The second one does not have the internal argument and it denotes the event itself (i.e., the action or process of innovating).

of the events as exemplified in the sentence below. Such nominalizations are labeled as NN, as they can no longer be regarded as event nominals.

(1) 这 项 发明 很 重要.
zhè xiàng fāmíng hěn zhòngyào.
this CL innovation very important
"This piece of innovation is very important"

Syntactic analysis above also offers a novel perspective on the arguments of event nominals. It suggests that arguments are all introduced by functional projections as specifiers (Ramchand 2008). Specifically, for typical verbs or fact-reading nominal events, the vP projection categorizes a category-neutral root and introduces the internal argument which is the entity undergoing change (*undergoer*); and the VoiceP specifies a causation subevent and introduces the external argument which initiates the event

(*initiator*). For event-reading ones, their syntax-driven arguments solely consist of the *undergoer* (albeit as a modifier). The semantic roles of other so-called arguments are inferred subjectively due to the absence of other event-related projections at the syntactic level.

3. Identification of Event Nominals

We conduct a manual annotation and an automatic tagging experiment to investigate the effectiveness of the UG-inspired schema in identifying nominal events of Mandarin Chinese. Annotators need to identify event nominals before and after acquiring the above classification and its underlying principles, which could be perceived as a deductive annotation guideline. Equipped with such predefined criteria or rules, they are thought to be able to analyze various cases and make decisions accordingly, without the need for an extensive enumeration of diverse phenomena. Furthermore, we leverage the UG-inspired annotation materials to train automatic taggers, subsequently comparing their performance against those trained on an existing resource where event nominals are tagged without the assistance of the UG-inspired schema.

We investigate that the incorporation of the UG-inspired schema yields higher IAA among human annotators, along with accuracies/F_1 scores of higher level exhibited by automatic taggers when applied to our newly established dataset. These outcomes are considered indicative of the validity and efficacy of the UG-inspired classification.

3.1 Manual Tagging

3.1.1 Data Source. A total of 1,300 sentences extracted from Chinese TreeBank (CTB; Xue et al. 2005), where text from *People's Daily* is fully segmented and POS tagged, are annotated in total. However, CTB only annotates predicative verbs as VV, while all the other non-typical ones are labelled NN. Without a dedicated tag for nominalizations, we cannot distinguish them from typical nouns and, hence, an external source, Corpus of People's Daily (Yu, Duan, and Wu 2018), is utilized considering its usage of a VN tag as well as its similar focus on newspaper articles. Tokens tagged VN in it constitute a seed lexicon where all its members have the potential to be nominalized. When they appear in CTB sentences and are tagged as verbs or nouns, they are selected as our annotation targets.

3.1.2 Procedure. Initially, two doctoral students and one undergraduate student, native speakers of Mandarin and majors in linguistics, are required to differentiate verbal and nominal events, following the manner of CTB in handling event nominals. Later, they acquire the UG-inspired schema as well as its rationale. Under its guidance, they are required to differentiate the three types of nominalizations, and assign VV, VN, and NN tags accordingly in the new data. After some iterations in comparing and discussing their disagreements to ensure a correct understanding, the data left are annotated by a single annotator separately to boost efficiency of annotations.

3.1.3 Annotation Quality. Quality of our newly established corpus can be firstly gauged by IAAs among annotators. Specifically, for annotation without the assistance of the UG-inspired schema, POS tags, VV, and NN provided by CTB are used as a reference to divide our targets into verbal and nominal events. IAAs on each category are then calculated respectively. For annotation guided by the schema, there is no gold-standard annotation so calculation of IAAs in classifying each category of nominalizations is

Table 3
IAAs among annotators in the nominalization identification task with and without the guidance of the UG-inspired schema.

	VV	VN	NN	Overall
IAA − UG schema	94.67%	82.61%		88.02%
IAA + UG schema	97.13%	94.16%	87.85%	94.99%

based on the formula below. Calculating the overall IAAs for both of them is the same. The results are shown in Table 3.

$$IAA_{\text{categorial}-\text{UG schema}} = \frac{\Sigma_{\text{Agreed (VV, NN)}}}{\Sigma_{\text{(VV, NN)}}}$$

$$IAA_{\text{categorial}+\text{UG schema}} = \frac{\Sigma_{\text{Agreed (VV, VN, NN)}}}{\Sigma_{\text{Assigned by at least one annotator (VV, VN, NN)}}}$$

$$IAA_{\text{overall}} = \frac{\Sigma_{\text{Agreed Recordings}}}{\Sigma_{\text{Recordings}}}$$

Furthermore, we introduce an additional metric, entropy, which serves as an indicator of the level of uncertainty or informativeness, to assess the validity of our annotations inspired by UG. Computations of entropy values are either based solely on tag distribution or involve tokens together with their corresponding tags. The following formulas elucidate the specific methodology employed for the calculation of tag-level and token-level entropy values.

$$H_{\text{tag-level}} = -\sum_{i=1}^{m} p(\text{TAG}_i)\log_2(p(\text{TAG}_i))$$

$$H_{\text{token-level}} = -\sum_{i=1}^{n} p((\text{TOKEN, TAG})_i)\log_2(p((\text{TOKEN, TAG})_i))$$

As shown by Table 4, the higher entropy values in both tag-level and token-level of our data suggest it carries more information content compared with CTB data.

We argue that results of this annotation experiment, which are more consistent and more informative annotations, underscore the feasibility and validity of implementing the UG-inspired schema in the manual identification of event nominals within Mandarin Chinese.

Table 4
Overall entropy values, both in tag-level and token-level, of original CTB annotations and our newly created corpus.

	Tag-level Entropy	Token-level Entropy
CTB Data	0.9981	7.8461
Our Data	1.4220	7.9245

Table 5
Data distribution of our annotated corpus for nominalizations in Mandarin Chinese. #sentence, #VV, #VN, and #NN are the numbers of sentence, VV, VN, and NN, respectively.

	#sentence	#VV	#VN	#NN	Overall
Single-annotated	700	1,214	1,153	270	2,637
Triple-annotated	600	984	865	234	2,213
Total	1,300	2,198	2,018	504	4,850

3.1.4 Result. We thus introduce a new moderate-sized corpus containing high-quality manual annotations for nominalizations in Mandarin Chinese, whose data distribution is shown in Table 5. The data is now available at `https://github.com/MandarinMeaningBank/EventNominals`.

3.2 Automatic Tagging

We further conduct an experiment to test if models trained on our dataset are more capable of identifying nominalizations in Mandarin or not.

3.2.1 Experimental Setup

Data Preparation. Automatically detecting and classifying events can be viewed as a sequence labelling task, which requires all the tokens in the context as input. Therefore, we assign NULL labels for those that are irrelevant to our research target. All the annotation data is randomly split into training, development, and test sets, which hold 80%, 10%, and 10% instances, respectively. Model selection is judged by its performance on the development set while the reported tagging performance is obtained by applying the selected model on the test set.

Model. For this sequence labelling task, models need to receive a sequence of tokens $(w_1, w_2, \ldots, w_n)$ as input and automatically output labels (VV, VN, NN, and NULL) for each of them. Given the Bidirectional Long Short-Term Memory (BiLSTM) model, which is the bidirectional variation of the Long Short-Term Memory model (LSTM; Hochreiter and Schmidhuber 1997), has been proved effective and competitive in various sequential tagging tasks (Huang, Xu, and Yu 2015; Ma and Hovy 2016; Bohnet et al. 2018), we use it to build the baseline system for our data set. Input of our model is the embeddings of words $\boldsymbol{w}_i$ which have a dimension of 300 and are initialized by pre-trained word vectors (Li et al. 2018).[1]

Contextual word representations $\boldsymbol{c}_{w_i}$, which are output of the last layer of a pre-trained language model provided by the transformer library, bert-base-chinese,[2] are also leveraged as input of the BiLSTM model. Considering the model is pretrained on large

1 `https://github.com/Embedding/Chinese-Word-Vectors`.
2 `https://huggingface.co/bert-base-chinese`.

Table 6
Average auto-tagging accuracies/F_1 scores on each category and the entire dataset (either including NULL or not) of different models (from 5 runs). In assessing the overall performance across various categories, the weighted method, which assigns weights to individual classes based on their quantity of samples, is used. Overall (+/−NULL) denotes whether the calculation of accuracies and F_1 scores involves samples assigned to the label NULL or not.

	VV	VN	NN	Overall (−NULL)	Overall (+NULL)
BiLSTM	84.31/88.57	82.99/85.98	69.81/68.75	82.51/85.30	97.31/99.22
BiLSTM+CRF	86.42/89.87	85.92/81.16	71.93/76.19	84.98/84.62	97.67/98.85
Bert	92.91/88.37	93.62/86.36	78.17/83.33	92.01/86.67	98.06/99.27
Bert+CRF	95.14/93.02	94.76/91.30	79.53/80.00	93.66/90.91	98.79/99.71

corpus with character-level vocabulary, the out-of-vocabulary issue could be addressed to an extent.

$$\overrightarrow{\boldsymbol{h}}_i = \overrightarrow{\text{LSTM}}\left(\overrightarrow{\boldsymbol{h}}_{i-1}, \left[\boldsymbol{w}_i; \boldsymbol{c}_{w_i}\right]\right)$$
$$\overleftarrow{\boldsymbol{h}}_i = \overleftarrow{\text{LSTM}}\left(\overleftarrow{\boldsymbol{h}}_{i+1}, \left[\boldsymbol{w}_i; \boldsymbol{c}_{w_i}\right]\right)$$

The output of BiLSTM $\boldsymbol{h}_i$, which is the concatenation of $\overrightarrow{\boldsymbol{h}_i}$ and $\overleftarrow{\boldsymbol{h}_i}$, is then fed into a softmax layer or a Conditional Random Field (CRF) layer to infer tags for each token independently or a sequence of tags. Parameters of our models are given in Appendix A.

3.2.2 Main Results. Table 6 shows performance of different models in identifying each category. With the exception of NN, whose performance is largely hindered by sparse data, by adding a BERT and CRF layer, other event-nominals-associated categories, VV and VN, could achieve high-level performance, 95.14/93.02 and 94.76/91.30 accuracies or F_1 scores, respectively. The overall performance also exhibits a notable level. In our view, all these indicate the deductive guideline hereby proposed responds in a considerable measure to the concerns on identifying and classifying event nominals.

3.2.3 Comparative Analysis. To further substantiate the validity and the robustness of our annotations within the UG-inspired schema, a comparative analysis between it and an existing annotation resource, the Corpus of People's Daily (Yu, Duan, and Wu 2018) wherein nominal events are also annotated, is conducted. More specifically, VN and NULL instances, with a data distribution closely resembling that of our newly annotated dataset, are drawn from the Corpus of People's Daily, serving as a reference dataset.[3] It is then subjected to the same experimental setup in §3.2.2. We contend that comparing performance metrics between the two experiments is plausible and equitable as (i) both annotation resources originate from articles published in the *People's Daily* newspaper, albeit from different years; (ii) and a similar data distribution of the two datasets is maintained.

3 The reference dataset contains 2,018 VN instances and 31,645 NULL instances.

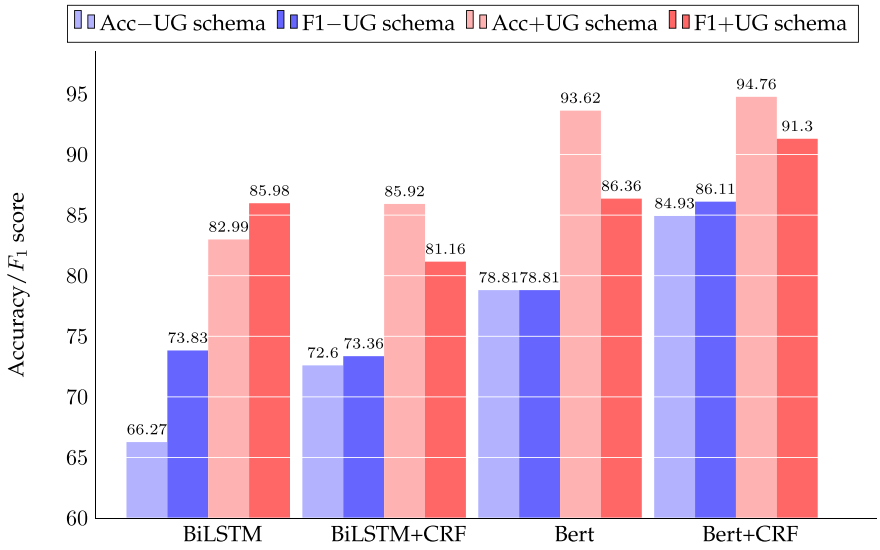

Figure 7
Tagging accuracies and F_1 scores on annotations +/− UG-inspired schema. Metrics of + UG-inspired schema are consistent with the results of identifying VN in the prior experiment.

Figure 7 provides a comprehensive exposition of the experimental outcomes derived from these corpora. It can be seen that our annotations, guided by the UG-inspired schema, facilitate a more accurate and expeditious identification of nominal events by automated taggers.

An ablation experiment is then conducted to assess the influence of the size of our corpus made on the performance. As noted in Figure 8, a considerably high accuracy

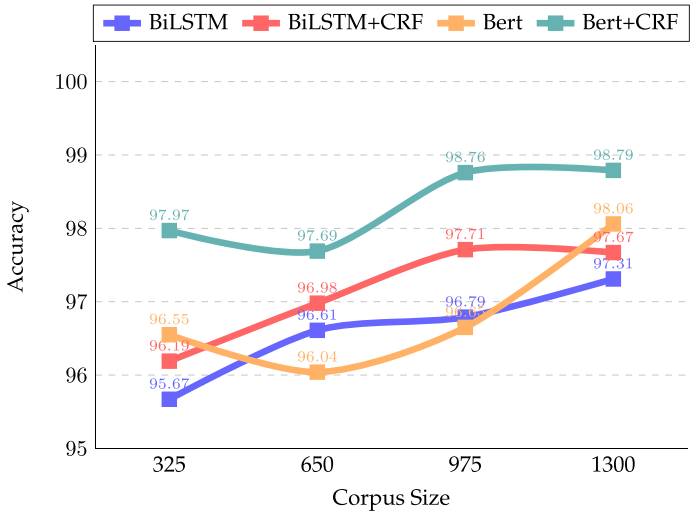

Figure 8
Auto-tagging accuracies of different models based on various sizes of our newly-created corpus. Accuracies are calculated by taking all samples (including NULL) into account.

could be achieved even with a relatively small dataset. We reaffirm the validity, as well as the reliability, of our annotations.

4. Nominal SRL

As mentioned before, arguments of event nominals in different languages may involve complex interactions between syntactic and semantic properties, and thus pose a great challenge in labelling them in a consistent and crosslingually compatible way.

Faced with this notorious difficulty, we argue that the UG-inspired schema could contribute to the nominal SRL task by providing annotators with great clarity on and more precise knowledge of non-adverbial arguments rather than treating them as prototype agents or prototype patients. Within this schema, both internal and external arguments or, in other words, **undergoers** and **initiators**, could be purely defined through their syntactic properties.

Moreover, for event nominals with event reading[4] in Mandarin Chinese, analysis inspired by the schema suggests we could classify their arguments to two types: Contrasting with the pattern "internal argument+event-reading event nominals," or more concisely, "ARG+VN", which is initially fixed by syntactic structures (see Figure 6), other so-called arguments in the patterns like "ARG+DE+VN" and 'VN+ARG" are identified through semantic or pragmatic criteria, which require annotators to infer their semantic roles with events according to semantically variable contexts. The following are some illustrative examples of frequent patterns in our annotations, and the remaining cases that cannot be categorized into the three types are labelled as OTHER.

- ARG+VN

(2) 城市 建设
chéngshì jiànshè
city construction
"the city construction"

- ARG+DE+VN

(3) 城市 的 建设
chéngshì de jiànshè
city DE construction
"the city construction"

4 For fact-reading event nominals, they project arguments just like typical verbs (see Figure 5) which could be temporarily overridden; and for entity-reading nominalizations, they do not have arguments. Therefore, we will solely focus on event-reading nominal events hereafter.

- VN+ARG

(4) 建设 人员
jiànshè rényuán
construction people
"people who construct (sth)"

Three annotation experiments, each of which present 1,000 event nominals along with candidates for their arguments,[5] are conducted to test if the definition and classification of arguments grounded on the UG-inspired schema could be supported by empirical evidence.

4.1 Baseline Experiment

Prior to formal experiments, we conduct a minimal annotation task as baseline to establish a reference point or benchmark of IAAs. To make a reasonable comparison, all the variables are held constant with those in the following formal ones.[6] Details of the annotation process include:

- following specifications of Chinese NomBank, annotators are required to differentiate ARG0 (proto-Agent) and ARG1 (proto-Patient)
- if they cannot determine which labels should be assigned, the tokens will be marked as "hard case" on an experimental basis
- regarding "hard case" as a penumbra of possibilities, which could be annotated consistently or not by chance, the observed proportionate agreement is calculated in two versions

$$IAA_{\text{+hard case}} = \frac{\Sigma_{\text{agreement}}}{\Sigma_{\text{record}}}$$

$$IAA_{\text{–hard case}} = \frac{\Sigma_{\text{agreement}}}{\Sigma_{\text{record}} - \Sigma_{\text{hard case}}}$$

Table 7 shows the result of this baseline experiment, which could be regarded as the starting point or initial conditions of our target tasks.

5 All the annotation data are originally from the CTB, and are split to three parts based on IDs. In this way, a degree of equilibrium in data distribution across the three successive nominal SRL tasks is guaranteed and IAAs observed are thus amenable to comparison.

6 The contents of the materials, which are annotation outcomes of §3, are not the same in the three experiments, despite all the other variables being identical. Such an avoidance of data reusing in the three experiments is to alleviate plausible concern that an elevation in IAAs potentially stems from annotators' familiarity and adeptness with the content. Also, though annotators have previously encountered the data in the identification task, we contend that the utilization of annotations of §3 is not only necessary (the prerequisite for the nominal SRL task is to have suitable candidates for event nominals) but also may not pose issues, as the two tasks are inherently different from each other.

Table 7
IAAs of our annotators following the guideline of Chinese NomBank. +/− hard case means whether they are taken into account or not.

	+hard case	–hard case
IAAs in Baseline Experiment	81.31%	90.46%

4.2 Experiment 1: The Role of Syntax

The first experiment aims to assess if the introduction of a more clear-cut definition of non-adverbial arguments provided by the UG-inspired schema could develop the common understanding of annotators, which naturally leads to improvements in IAAs.

Materials. We extract 1,000 phrases according to our former annotations towards event nominals. Each phrase contains both a token tagged VN and at least one token tagged as arguments by Chinese NomBank. Grammar trees provided by CTB are utilized as a reference source to pick out sequences that match this criteria.

Participants. Three native speakers of Mandarin Chinese participate in the experiment. Both of them major in linguistics and are taught the classification and its underlying rationales in §2.

4.3 Experiment 2: The Role of Context

As previously hypothesized, there are two types of arguments of event-reading event nominals: arguments identified and labelled according to formalized criteria and those whose semantic roles are inferred based on contexts or world knowledge of annotators. Experiment 2 thus tests if and how the contextual information influences the annotation task.

Materials. We newly extract 1,000 cases. Compared with the extracted phrases in §4.2, these cases include all the tokens from sentences, and thus could be considered to convey more contextual information when other conditions are kept the same.

Participants. Participants in the experiment are the same as those in §4.2.

4.4 Results and Analysis

Table 8 presents the results of these three experiments, which indicate that the IAA in Experiment 1, with the assistance of the purely syntactic definition, is increased by 3.15% (90.46% vs. 93.75%, $p < 0.05$ according to one-way ANOVA) compared with the annotation under the guideline of Chinese NomBank. Furthermore, the IAA in Experiment 2 also increases by 7.58% (98.04% vs. 90.46%, $p < 0.01$) compared with the initial annotation, and, most importantly, the IAA in this experiment is also significantly higher than that in Experiment 1 (98.04% vs.93.75%, $p < 0.01$). Therefore, we can assert that the syntactically formalized criteria, as well as the context information, indeed contribute to improving IAAs in the nominal SRL task.

Table 8
IAAs between our annotators in three experiments. +/– hard case means whether they are taken into account or not.

	+hard case	–hard case
IAAs in Baseline Experiment	81.31%	90.46%
IAAs in Experiment 1	89.57%	93.75%
IAAs in Experiment 2	94.09%	98.04%

To further examine which annotation targets, syntax and context, bring greater positive influence, we also leverage a fine-grained analysis towards each syntactic pattern. Figure 9 shows IAAs under each of them in these three successive annotation experiments.

We compare the IAA increases of different patterns in two directions to obtain more insights.

- Horizontal: The relationship between IAA increase of each pattern with the overall IAA increase is fitted within a *multiple linear regression* model. After regression modeling, the regression coefficients are tested for significance. In this way, we can identify which pattern would have a significantly positive impact on the overall IAA improvement.
- Vertical: Differences between IAAs within one pattern is compared using the one-way ANOVA method. Multiple comparisons using Bonferroni correction further indicate between which two rounds of annotations

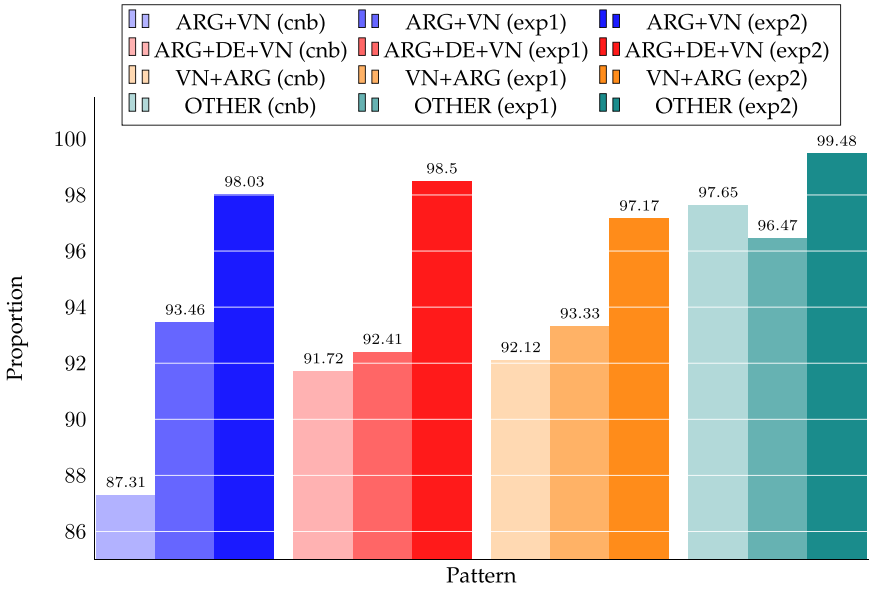

Figure 9
IAAs of different patterns in three experiments. cnb refers to Chinese NomBank; exp1 refers to experiment 1 and exp2 refers to experiment 2.

Table 9
Horizontal and vertical comparison of IAA differences of patterns in the first two annotation experiments.

	horizontal	vertical
ARG+VN	$p = 0.006^{**}$	$p = 0.001^{**}$
ARG+DE+VN	$p = 0.799$	$p = 0.936$
VN+ARG	$p = 0.554$	$p = 0.884$
OTHER	$p = 0.208$	$p = 0.195$

$^{**}p<0.01$.

such differences exist. The result, namely, whether they differ from each other significantly, could reveal the pattern for which the IAA experiences significant improvement upon the application of syntactic theory or contextual information.

Comparisons between Chinese-NomBank-style annotations and those of Experiment 1 are shown in Table 9. Based on this, we conclude the growth of the overall IAA in Experiment 1 is mainly due to the pattern ARG+VN ($p < 0.01$), and also, it is the only pattern whose two IAAs differ significantly (87.31% vs. 93.46%, $p < 0.01$). Both of these highlight the importance of a formalized criterion in annotating nominal SRL: It benefits the annotation task by providing a clear-cut line between ARG0 and ARG1 and this influence is particularly evident in the syntactically fixed pattern ARG+VN.

Turning to Experiment 1 and Experiment 2, comparison results are detailed in Table 10. According to statistics, it is evident that while IAAs of all the patterns demonstrate improvement, only those of the ARG+VN and ARG+DE+VN patterns exhibit statistically significant enhancements in overall IAA. Furthermore, the incorporation of more contextual information also notably impacts the IAA improvements of these two patterns.

After detailed investigation, we are of the opinion that the reason behind the significant impact of context on the two above patterns is that annotators cannot differentiate argument types within local distance. However, if we do not limit context to additional tokens we newly introduce, the other two patterns could also be regarded to benefit from the context information: The VN+ARG pattern is similar to predicative verbs with objects in forms, which could assist annotation by providing a short-length reference pattern; and the OTHER pattern always includes ARG0 and ARG1 tokens within itself,

Table 10
Horizontal and vertical comparison of IAA differences of patterns in the last two experiments.

	horizontal	vertical
ARG+VN	$p = 0.001^{**}$	$p = 0.022^{*}$
ARG+DE+VN	$p = 0.032^{*}$	$p = 0.001^{**}$
VN+ARG	$p = 0.717$	$p = 0.260$
OTHER	$p = 0.681$	$p = 0.164$

$^{**}p<0.01$; $^{*}p<0.05$.

which makes them easily identifiable. The only difference between them and those patterns with significantly improved IAAs is whether the context is local or not.

Supported by our experimental results, we conclude that it is reasonable to draw a line between ARG+VN and other types of patterns as defined by the UG-inspired linguistic analysis towards arguments. The distinction, from our point of view, could benefit the nominal SRL task in two ways. Firstly, it could predict possible difficulties in annotations. For which annotation target can annotators come to agreement easily or even approach 100% IAAs and for which should we not anticipate satisfactory results even after several iterations of re-tagging? In Experiment 1, the contrast between the significantly increased IAA of ARG+VN and those of others is a clear example. What is more, it could also provide clues on how to remove obstacles that stand in the way of the continuing improvement on IAAs for some cases—both syntax and context information are proved to be valid in advancing overall annotation quality in the two experiments.

5. Discussion

5.1 Expert vs. Crowdsourced Annotation

Annotations have a significant influence on almost all stages of machine learning (Ide and Pustejovsky 2017), not only in terms of scale but also quality, as it is intuitively assumed that the annotation quality (always approximated by the IAA), determines the upper limit of models' performance (Gale, Church, and Yarowsky 1992; Navigli 2006; Bender et al. 2015).[7] Given that, recent years have witnessed a growing interest in annotation quality control in the NLP community (Snow et al. 2008; Cerezo, Bravo-Marquez, and Bergel 2021; Noh et al. 2020; Hahn et al. 2012) to lay a solid foundation for further optimizing metrics of systems. Nonetheless, we notice that most existing studies only take extrinsic influencing factors such as annotators, annotation procedures, or tools into account, which may help reduce accidental errors but are useless in so-called hard cases—there exist some situations whose unsatisfying qualities stem from the ill-defined tasks themselves. In our view, event nominals in multiple languages are such a case that confuses not only annotators or machines but also guideline-makers because of the lack of formalized and consistent criteria.

From this perspective, the attempt in this article, which is inspired by UG and maps insights of event nominals drawn from inflectional languages to non-inflectional Mandarin Chinese, could also be viewed as an exploration on how an intrinsically hard annotation target could be mitigated or even solved. We demonstrate that with the support and guidance of post-traditional theoretical linguistics, our professional annotators can be inspired by typologically distant languages, which could be thought of as a genuine solution of the notorious difficulty.

Our expert annotation also implies that there are important considerations in utilizing either the expert or the crowdsourcing method in practice. The expert annotation inherently carries a limitation that it is proactive and demanding to generate large-scale data, as evidenced by the size of our newly established corpus in this work. In light of this, it is imperative to scrutinize the impact of corpus size on auto-tagging performance, thus shedding light on whether models trained on a relatively small dataset can attain a promising level of performance. For crowdsourcing annotations,

7 There are systems outperforming IAAs. Boguslav and Cohen (2017) point out that these counterexamples are not random noise but rather reflect a real phenomenon.

noting simplifying tasks is always necessary in order to meet backgrounds of non-expert annotators (He, Lewis, and Zettlemoyer 2015; Michael et al. 2018; Klein et al. 2020); this article stresses that omitting information should be handled with extreme care—simplification on morpho-syntactic heterogeneity of event nominals may be the most expeditious but not the most effective approach.

5.2 Verbal vs. Nominal Events

Events play a central role in natural language understanding, not only as verbal cores of sentence-level meaning by associating participants, namely, who does what to whom and when/where the event takes place, but also as objective individuals that could be referred, namely, nominal events. Both of them overwhelmingly contribute to the information required by a series of downstream tasks (Yang et al. 2003; Glavas and Najder 2014). Nonetheless, compared with verbal events (Gildea and Jurafsky 2002; Kinyon and Prolo 2002; Srikumar and Roth 2011; Judea and Strube 2017), development of studies targeting nominal ones in the NLP community is somewhat limited. In this regard, the present study could be viewed as an endeavor to conduct a deeper investigation into the linguistic phenomenon.

We posit that such an exploration carries significance due to the substantial role that nominal events play—in our view, nominalizations enable humans to refer to events or attribute their properties, and then facilitate discussing them. It is thus conceivable that this communicative act is both universally available across languages and frequently utilized by language users. Our statistical finding that 47.86% of identified events are categorized as nominal instances further supports this point.

Our study also yields insights on events by introducing new-stage linguistic theories. Within the Minimalist Approach, functional items are hierarchically organized around events by viewing, anchoring, or linking them. This perspective affords us the opportunity to conduct an in-depth analysis of events, encompassing their internal temporal structures and their intricate interplay with the external time (Reichenbach 1947; Vendler 1967; Davidson 1969). We contend that these insights bear the potential to enhance various event-centric tasks and their corresponding methodologies. Specifically, they hold relevance for event detection (Chen et al. 2015; Huang et al. 2018) and the establishment of event-event relationships, especially temporal (Ning, Wu, and Roth 2018; Ning et al. 2018; Han, Ning, and Peng 2019), co-referential (Nothman et al. 2012; Peng, Song, and Roth 2016; Barhom et al. 2019), and hierarchical relations (Araki et al. 2014; Badgett and Huang 2016; Aldawsari and Finlayson 2019).

5.3 Syntax-based vs. Situation-based SRL

Towards the SRL task or more precisely, SRL of verbs, which is regarded as the backbone of a sentence, researchers have developed a range of lexical resources, including FrameNet (Baker, Fillmore, and Lowe 1998), PropBank (Kingsbury and Palmer 2002), VerbNet (Schuler 2006), and so forth. They can be classified into two types given that linguistic theories relied upon by them have different criteria or basis in identifying arguments.[8]

8 https://natural-language-understanding.fandom.com/wiki/Semantic_role_labeling.

- Syntax-based: This approach is represented by VerbNet (Schuler 2006). It is an extension of Levin's (1993) classification, whose basis is distinguishable syntactic forms of verbs and their arguments. Arguments under this approach are sentence-specific or construction-specific, and must be explained syntactically.
- Situation-based: This approach is represented by FrameNet (Baker, Fillmore, and Lowe 1998) whose theoretical background is Frame Semantics (Fillmore 1965, 2006). It is asserted that a word, regardless of its syntactic category, can invoke a semantic frame with multiple frame-specific elements. The classification of frames, and consequently, sets of possible frame elements, are totally dependent on encyclopedic knowledge.

Turning to SRL of nominal events from that of events triggered by predicative verbs, to the best of our knowledge, existing frameworks and annotation resources, such as NomBank (Meyers et al. 2004a, 2004b), QANom (Klein et al. 2020), and Chinese NomBank (Xue 2006), all solely use the situation-based approach. They follow the line of Frame Semantics (Fillmore and Atkins 1994; Fillmore and Baker 2001) and contend that corresponding verbal and nominal events activate the same situation with the same organization of semantic roles. Therefore, whether in monolingual or multilingual settings, annotators can identify event nominals and perform nominal SRL by comparing them with their predicative counterparts without considering their different surface forms.

Nonetheless, as evidenced by the less satisfactory performance associated with nominal events compared with verbal ones shown in Table 11, we argue that this approach is less than ideal. While it facilitates annotations and allows for extension to multiple languages, the approach heavily relies on annotators' vulnerable subjective judgment, which can vary based on personal experience and worldviews. As a result, it may lead to inconsistent and unstable annotations. This article thus adopts the syntax-based perspective towards nominal SRL. Under a purely syntactic framework, the articulation of the complex and mixed concept *argument* results in greatly improved IAAs.

The distribution of argument types in our annotation also indicates the vital and indispensable role of our attempt, the application of a syntax-based way in nominal SRL. For the pattern ARG+VN, though its meaning is the same as that of ARG+DE+VN, their proportions in argument types, as shown by Table 12, are totally different from each other (17.79% + 82.21% vs. 44.54% + 55.46%). Such a difference, from our perspective, is rooted in their forms—the appearance of DE relaxes fixed syntactic pattern and incorporates more ARG0 unconditionally. Unbalanced distributions of argument types in VN+ARG pattern (62.99% vs. 37.01%) is another evidence. We believe it is the similarity between VN+ARG1 and predicative verbs with their objects in forms that restricts its use.

All these re-emphasize that we should probably not push denotations of all the expressions, with which the world is voiced, to the mushy situation-based meaning, but recognize the influence brought by syntax, or more generally forms, to meaning.

6. Conclusion

From predicate-argument structures to more general event structures, our introduction of a new-stage Chomskyan approach, which is the focus of more than a decade of

Table 11
With the situation-based approach, "accuracies" of verbal/nominal events identification and SRL tasks for both human and machines in English and Mandarin Chinese. For verbal events, the IAA on identification task in English is inferred according to the annotation of Fan et al. (2011); the accuracy is inferred by the percentage false predictions of the Stanford tagger reported by He, Lewis, and Zettlemoyer (2015); the IAA on identifying those of Mandarin is from our own annotation; the performance of machines is approximated by F_1 scores of two subcategories of verbs, VA and VC reported in Sun and Wan (2016); and the data of SRL are provided by the PropBank (Kingsbury and Palmer 2002) and the Chinese PropBank (Xue and Palmer 2003), respectively. For nominal events, QANom project (Klein et al. 2020) reports the IAAs and F_1 scores on identification task in English; the IAA on identifying those of Mandarin comes from our own annotation; the performance of machines is evaluated by F_1 scores in Zhao et al. (2007); IAAs on English nominal SRL are provided by NomBank (Meyers et al. 2004b, 2004a); Liu and Ng (2007) report the the F_1 score of this task; IAAs on Mandarin nominal SRL are from our own annotation experiment on a role classification task towards non-adverbial arguments following the guideline of Chinese NomBank (Xue 2006) and Li et al. (2009) reports the F_1 score based on the data of Chinese NomBank (Xue 2006).

	Verbal Event				Nominal Event			
	Identification		SRL		Identification		SRL	
	English	Mandarin	English	Mandarin	English	Mandarin	English	Mandarin
Human	∼98%	88.58%	∼90%	92.5%	81.8%–85.6%	82.61%	∼85%	84.8%
Machine	97%–98.5%	81.47%–96.01%	∼93%	94.1%	82.6%	84.1%	77.0%	72.7%

Table 12
Distributions of argument types of event nominals in Mandarin Chinese, based on our annotation data.

	ARG+VN	ARG+DE+VN	VN+ARG
ARG0	17.79%	44.54%	62.99%
ARG1	82.21%	55.46%	37.01%

intensive study in theoretical linguistics but long-overlooked by computational linguistics, brings new insights to language universals. Inspired by the universals it describes, we demonstrate how challenges in handling event nominals and arguments across languages, that is, their surface-level morpho-syntactic heterogeneity, could be identified and addressed. By assuming that universals are abstract functions of functional constituents, it is feasible to uniformly define event nominals in different languages as various combinations of event- or entity-related functional projections. This encompassing framework can also extend to a non-inflectional language, Mandarin Chinese.

In our view, this UG-inspired schema, along with annotations within its guidance, contribute to both theoretical and computation linguistics. Theoretically speaking, the schema aligns with a research trajectory termed Generative Typology (Roberts 1997; Baker 2012), which shares common ground with traditional Functionalist Typology (Greenberg 1963) in pursuit of identifying shared patterns in representative samples from all natural languages but diverges from it at the level of theoretical abstraction. By abstracting UG, a generative notion from the functions of functional constituents in typologically distant languages, our article could thus be viewed as an endeavor striking a balance between language diversity and universality, and thus contributing to the ongoing discussion surrounding UG (cf. Evans and Levinson 2009). From an

empirical perspective, the UG-inspired schema has been demonstrated to help the resolution of the event nominals identification and nominal SRL tasks by establishing a clear understanding of how event nominals and their arguments at different levels can be made to fit together in a comprehensive, coherent, and effective architecture. It not only provides us with a deductive, generative, and clear-cut specification—which is proved to be compatible for nominalizations in Mandarin and therefore enables us to achieve relatively high IAAs and auto-tagging accuracies—but also suggests that arguments of event nominals can be divided into two types: one is syntax-driven and can be more easily accessed by involving linguistic theories, while the other is on a semantic or a pragmatic basis and its annotation quality can be improved by incorporating more context information.

It is also imperative to recognize that the schema has only demonstrated its compatibility with Mandarin in this work. Extending it to other languages requires additional scrutiny. From a broader perspective, we perceive this as an invitation for expanded inquiry. We expect our successful experience, characterized by the incorporation of post-1990s linguistic theories to systematically reexamine linguistic universals, could offer motivation as well as a valuable reference for researchers in the field of NLP to embark on more in-depth investigations encompassing a diverse array of languages. We believe such endeavors hold the promise of yielding insights of greater depth and breadth, thereby augmenting the efficacy of more specific NLP tasks.

Appendix A. Implementation Details of Our Models

Details of parameter settings in our models are as follows: (i) all the models are trained with a batch size of 64; (ii) the dimension of the hidden states of BiLSTM is set to 128 for each direction and the number of layers is set to 1; (iii) for our baseline system, learning rates are chosen from $\{1e-4, 5e-3, 1e-3\}$, while for models involving pretrained language models, selection of learning rates are in accordance with those in the original paper, which are $\{2e-5, 3e-5, 5e-5\}$.

Intuitively, the CRF layer performs better as orders of tags are useful information in sequence tagging tasks and should be taken into account. We also notice that initiating CRF transitions conditionally and setting learning rates respectively for BiLSTM and CRF can increase accuracies—they can be increased from 98.18% to 98.79% after using these two tricks.

All the models are trained on one A100 GPU.

Acknowledgments

We express our gratitude for the assistance and support rendered by annotators Nan Li, Xihao Wang, and Chunhui Sun. We extend special appreciation to Simone Teufel for her insightful suggestions. Additionally, our heartfelt thanks go to all the reviewers and editors for their significant contributions.

References

Abend, Omri and Ari Rappoport. 2013. UCCA: A semantics-based grammatical annotation scheme. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 1–12.

Abzianidze, Lasha and Johan Bos. 2017. Towards universal semantic tagging. In *IWCS 2017—12th International Conference on Computational Semantics—Short papers*, pages 1–9.

Aldawsari, Mohammed and Mark Finlayson. 2019. Detecting subevents using discourse and narrative features. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4780–4790. https://doi.org/10.18653/v1/P19-1471

Alexiadou, Artemis. 2020. D vs. n nominalizations within and across

languages. In Artemis Alexiadou and Borer Hagit, editors, *Nominalization: 50 Years on from Chomsky's Remarks*. Oxford University Press, pages 88–109. https://doi.org/10.1093/oso/9780198865544.001.0001

Alexiadou, Artemis, Gianina Iordăchioaia, Florian Schäfer, Petra Sleeman, and Harry Perridon. 2011. Scaling the variation in romance and Germanic nominalizations. In Petra Sleeman and Harry Perridon, editors, *The Noun Phrase in Romance and Germanic*. John Benjamins, pages 25–40. https://doi.org/10.1075/la.171.04ale

Araki, Jun, Zhengzhong Liu, Eduard Hovy, and Teruko Mitamura. 2014. Detecting subevent structure for event coreference resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4553–4558.

Badgett, Allison and Ruihong Huang. 2016. Extracting subevents via an effective two-phase approach. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 906–911. https://doi.org/10.18653/v1/D16-1088

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 17th International Conference on Computational Linguistics*, volume 1, pages 86–90. https://doi.org/10.3115/980451.980860

Baker, Mark C. 2012. *Formal Generative Typology*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199544004.013.0012

Barhom, Shany, Vered Shwartz, Alon Eirew, Michael Bugert, Nils Reimers, and Ido Dagan. 2019. Revisiting joint modeling of cross-document entity and event coreference resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4179–4189. https://doi.org/10.18653/v1/P19-1409

Bender, Emily M., Dan Flickinger, Stephan Oepen, Woodley Packard, and Ann Copestake. 2015. Layers of interpretation: On grammar and compositionality. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 239–249.

Boguslav, Mayla and Kevin Cohen. 2017. Inter-annotator agreement and the upper limit on machine performance: Evidence from biomedical natural language processing. *Studies in Health Technology and Informatics*, 245:298–302.

Bohnet, Bernd, Ryan McDonald, Gonçalo Simões, Daniel Andor, Emily Pitler, and Joshua Maynez. 2018. Morphosyntactic tagging with a meta-BiLSTM model over context sensitive token encodings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2642–2652. https://doi.org/10.18653/v1/P18-1246

Bresnan, J. 2001. *Lexical-Functional Syntax*. Blackwell, Malden, MA.

Cerezo, Jhonny, Felipe Bravo-Marquez, and Alexandre Henri Bergel. 2021. Tools impact on the quality of annotations for chat untangling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 215–220. https://doi.org/10.18653/v1/2021.acl-srw.22

Chen, Yubo, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–176. https://doi.org/10.3115/v1/P15-1017

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA. https://doi.org/10.21236/AD0616323

Chomsky, Noam. 1993. A minimalist program for linguistic theory. In Kenneth Hale and Samuel Jay Keyser, editors, *The View From Building 20: Essays in Linguistics in Honor of Sylvain Bromberger*. MIT Press, Cambridge, MA, pages 167–176.

Chomsky, Noam. 1995. *The Minimalist Program*. MIT Press, Cambridge, MA.

Chomsky, Noam. 2007. *Interfaces + Recursion = Language?* De Gruyter Mouton, Berlin, New York.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

Davidson, Donald. 1969. The individuation of events. In Nicholas Rescher, editor,

Essays in Honor of Carl G. Hempel. Reidel, Dordrecht, pages 216–234. https://doi.org/10.1007/978-94-017-1466-2_11

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Evans, Nicholas and Stephen C. Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5):429–448. https://doi.org/10.1017/S0140525X0999094X, PubMed: 19857320

Fan, Jung wei, Rashmi Prasad, Rommel M. Yabut, Richard M. Loomis, Daniel S. Zisook, John E. Mattison, and Yang Huang. 2011. Part-of-speech tagging for clinical text: Wall or bridge between institutions? *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2011:382–391.

Fillmore, Charles J. 1965. The case for case. In Emmon W. Bach and Robert Thomas Harms, editors, *Universals in Linguistic Theory*. Holt, Rinehart, and Winston, London, pages 1–25.

Fillmore, Charles J. 2006. Frame semantics. In Dirk Geeraerts, editor, *Cognitive Linguistics: Basic Readings*. Mouton de Gruyter Berlin, chapter 10, pages 373–400. https://doi.org/10.1515/9783110199901.373

Fillmore, Charles J. and B. T. S. Atkins. 1994. *Starting Where the Dictionaries Stop: The Challenge for Computational Lexicography*. Clarendon Press, Oxford, UK. https://doi.org/10.1093/oso/9780198239796.003.0013

Fillmore, Charles J. and Collin F. Baker. 2001. Frame semantics for text understanding. In *Proceedings of WordNet and Other Lexical Resources Workshop, NAACL*, volume 6, pages 59–64.

Gale, William, Kenneth Ward Church, and David Yarowsky. 1992. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *30th Annual Meeting of the Association for Computational Linguistics*, pages 249–256. https://doi.org/10.3115/981967.981999

Gildea, Daniel and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288. https://doi.org/10.1162/089120102760275983

Glavas, Goran and Jan Najder. 2014. Construction and evaluation of event graphs. *Natural Language Engineering*, 21:607–652. https://doi.org/10.1017/S1351324914000060

Greenberg, Joseph H. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Greenberg, Joseph H., editor, *Universals of Human Language*. MIT Press, Cambridge, MA, pages 73–113.

Hahn, Udo, Elena Beisswanger, Ekaterina Buyko, Erik Faessler, Jenny Traumüller, Susann Schröder, and Kerstin Hornbostel. 2012. Iterative refinement and quality checking of annotation guidelines—How to deal effectively with semantically sloppy named entity types, such as pathological phenomena. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3881–3885.

Han, Rujun, Qiang Ning, and Nanyun Peng. 2019. Joint event and temporal relation extraction with shared representations and structured prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 434–444. https://doi.org/10.18653/v1/D19-1041

He, Luheng, Mike Lewis, and Luke Zettlemoyer. 2015. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 643–653. https://doi.org/10.18653/v1/D15-1076

Hochreiter, Sepp and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computing*, 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735, PubMed: 9377276

Huang, Haoyang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2485–2494. https://doi.org/10.18653/v1/D19-1252

Huang, Lifu, Heng Ji, Kyunghyun Cho, Ido Dagan, Sebastian Riedel, and Clare Voss. 2018. Zero-shot transfer learning for event extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2160–2170. https://doi.org/10.18653/v1/P18-1201

Huang, Zhiheng, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Ide, Nancy and James Pustejovsky. 2017. *Handbook of Linguistic Annotation*, 1st edition. Springer Publishing Company. https://doi.org/10.1007/978-94-024-0881-2_1

Judea, Alex and Michael Strube. 2017. Event argument identification on dependency graphs with bidirectional LSTMs. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 822–831.

Kingsbury, Paul and Martha Palmer. 2002. From TreeBank to PropBank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, pages 1989–1993.

Kinyon, Alexandra and Carlos A. Prolo. 2002. Identifying verb arguments and their syntactic function in the Penn Treebank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, pages 1982–1988.

Klein, Ayal, Jonathan Mamou, Valentina Pyatkin, Daniela Stepanov, Hangfeng He, Dan Roth, Luke Zettlemoyer, and Ido Dagan. 2020. QANom: Question-answer driven SRL for nominalizations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3069–3083. https://doi.org/10.18653/v1/2020.coling-main.274

Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press.

Li, Junhui, Guodong Zhou, Hai Zhao, Qiaoming Zhu, and Peide Qian. 2009. Improving nominal SRL in Chinese language with verbal SRL information and automatic predicate recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1280–1288. https://doi.org/10.3115/1699648.1699674

Li, Shen, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on Chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 138–143. https://doi.org/10.18653/v1/P18-2023

Liu, Chang and Hwee Tou Ng. 2007. Learning predictive structures for semantic role labeling of NomBank. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 208–215.

Ma, Xuezhe and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074. https://doi.org/10.18653/v1/P16-1101

Meyers, Adam, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004a. Annotating noun argument structure for NomBank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 803–806.

Meyers, Adam, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004b. The NomBank project: An interim report. In *Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL 2004*, pages 24–31.

Michael, Julian, Gabriel Stanovsky, Luheng He, Ido Dagan, and Luke Zettlemoyer. 2018. Crowdsourcing question-answer meaning representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 560–568. https://doi.org/10.18653/v1/N18-2089

Navigli, Roberto. 2006. Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 105–112. https://doi.org/10.3115/1220175.1220189

Ning, Qiang, Hao Wu, and Dan Roth. 2018. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1318–1328. https://sdoi.org/10.18653/v1/P18-1122

Ning, Qiang, Ben Zhou, Zhili Feng, Haoruo Peng, and Dan Roth. 2018. CogCompTime: A tool for understanding time in natural

language. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 72–77. https://doi.org/10.18653/v1/D18-2013

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043.

Noh, Youngbin, Kuntae Kim, Minho Lee, Cheolhun Heo, Yongbin Jeong, Yoosung Jeong, Younggyun Hahm, Taehwan Oh, Hyonsu Choe, Seokwon Park, Jin-Dong Kim, and Key-Sun Choi. 2020. Enhancing quality of corpus annotation: Construction of the multi-layer corpus annotation and simplified validation of the corpus annotation. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 216–224.

Nothman, Joel, Matthew Honnibal, Ben Hachey, and James R. Curran. 2012. Event linking: Grounding event reference in a news archive. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–232.

Peng, Haoruo, Yangqiu Song, and Dan Roth. 2016. Event detection and co-reference with minimal supervision. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 392–402. https://doi.org/10.18653/v1/D16-1038

Ramchand, Gillian. 2008. *Verb Meaning and the Lexicon: A First-phase Syntax*. Cambridge University Press. https://doi.org/10.1017/CBO9780511486319

Ramchand, Gillian and Peter Svenonius. 2014. Deriving the functional hierarchy. *Language Sciences*, 46:152–174. https://doi.org/10.1016/j.langsci.2014.06.013

Reichenbach, Hans. 1947. *Elements of Symbolic Logic*. Macmillan & Co, New York.

Roberts, I. G. 1997. *Comparative Syntax*. A Hodder Arnold Publication. Arnold.

Schuler, Karin Kipper. 2006. VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon. Ph.D. thesis, University of Pennsylvania.

Snow, Rion, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast—But is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263. https://doi.org/10.3115/1613715.1613751

Srikumar, Vivek and Dan Roth. 2011. A joint model for extended semantic role labeling. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 129–139.

Sun, Weiwei and Xiaojun Wan. 2016. Towards accurate and efficient Chinese part-of-speech tagging. *Computational Linguistics*, 42(3):391–419. https://doi.org/10.1162/COLI_a_00253

Vendler, Zeno. 1967. Facts and events. In *Linguistics in Philosophy*. Cornell University Press. https://doi.org/10.7591/9781501743726

Wiltschko, Martina. 2014. *The Universal Structure of Categories: Towards a Formal Typology*, volume 142. Cambridge University Press. https://doi.org/10.1017/CBO9781139833899

Xue, Naiwen, Fei Xia, Fu-dong Chiou, and Marta Palmer. 2005. The Penn Chinese treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11(2):207–238. https://doi.org/10.1017/S135132490400364X

Xue, Nianwen. 2006. Annotating the predicate-argument structure of Chinese nominalizations. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 1382–1387.

Xue, Nianwen and Martha Palmer. 2003. Annotating the propositions in the Penn Chinese treebank. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, pages 47–54. https://doi.org/10.3115/1119250.1119257

Yang, Hui, Tat-Seng Chua, Shuguang Wang, and Chun-Keat Koh. 2003. Structured use of external knowledge for event-based open domain question answering. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and*

Development in Information Retrieval, SIGIR '03, pages 33–40. https://doi.org/10.1145/860435.860444

Yu, Shiwen, Huiming Duan, and Yunfang Wu. 2018. Corpus of Multi-level Processing for Modern Chinese. https://klcl.pku.edu.cn/gxzy/231686.htm.

Zhao, Jinglei, Changxiong Chen, Hui Liu, and Ruzhan Lu. 2007. Identification of Chinese verb nominalization using support vector machine. In *MICAI 2007: Advances in Artificial Intelligence*, pages 933–943. https://doi.org/10.1007/978-3-540-76631-5_89