Taiwan Journal of Linguistics
Vol. 13.1, 53-77, 2015
doi:10.6519/TJL.2015.13(1).3

AN INTEGRATED APPROACH TO FUNCTIONAL CORPUS CONSTRUCTION

Hengbin Yan[1] and Jonathan Webster[2]
[1]*Faculty of English Language and Culture, Guangdong University of Foreign Studies*
[2]*The Halliday Centre for Intelligent Applications of Language Studies, City University of Hong Kong*

ABSTRACT

In this paper, we present our recent experience in constructing a first-of-its-kind functional corpus based on the theoretical framework of Systemic Functional Linguistics. Annotated on selected texts from the Penn Treebank, the corpus was built by a collaborative team on a web-based annotation platform with several advanced features. After a discussion on the background and motivation of the project, we present our solutions to some of the challenges encountered in the collaborative annotation process. With fine-grained annotations of an initial corpus now available, the corpus can serve as a valuable linguistic resource that complements existing semantically annotated corpora and aids in the development of a larger-scale resource crucial for automated systems for analysis of linguistic function.

Key words: corpus annotation, linguistic function, collaborative annotation, functional semantics

1. INTRODUCTION

Recent years have seen data-driven approaches to natural language processing successfully applied to a wide range of problems including syntactic (Collins 2003; Klein and Manning 2003), semantic (Gildea and Jurafsky 2002; Pradhan et al. 2004) and discourse (Hernault et al. 2010) analysis. Computational processing of functional aspects of linguistic data, on the other hand, is a relatively underexplored research area. In linguistics, functional analysis refers to the study of language use in context. Among the theories for analyzing the functions of language, Systemic Functional Linguistics (SFL) (Halliday and Matthiessen 2004) is a linguistic framework that has become increasingly influential in recent years.

SFL is a theory about language use in context. Halliday argues that language is a social semiotic resource, a functional-semantic system. Functional-semantic in that language is functional, semantic, contextual (both textually and socio-culturally), and semiotic. The systemic functional approach to language is a systemic and a functional one, organized systematically into three metafunctions. Language is viewed in SFL as a semiotic device for making meaning. The way in which language expresses social and situational contexts and semantics is organized in networks of systems (sets of choices available at a certain point of context). Through realization rules the system networks are realized as functional structures and then as strings and sounds. Halliday argues that the system networks of meaning potential can be grouped into three broad functional categories, or strands of meanings: *ideational*, *interpersonal*, and *textual* metafunctions. Halliday discovers that interaction takes place in the systems within each of the metafunctions, but rarely across the metafunctions, thus making them independent of each other. As a semantic-oriented theory that focuses on language use in context - the role it plays in the social activity in which it is embedded, SFL regards language as primarily a resource for making meaning. It prioritizes meaning over form or rules, and is thus functional (what language means and how it is used) rather than formal (what language consists of). In other words, SFL researchers are generally more interested in how wordings and grammatical structures are used as a

means to construe meaning since they believe the actual form employed is less important than the function it performs.

SFL aims to explore how meanings are construed and take their current form in a text. The realizational analysis of SFL extends from the most abstract strata, for example, the context of ideology, culture and situation, to the most concrete realization of those meanings, for example, words, structures, phonology and graphology. This makes it possible to analyze texts both inductively and deductively. Thus, SFL provides an ideal handle for exploring language as intentional acts of meaning, complementing more syntactically oriented approaches to linguistic study. Despite its power, traditional analysis with SFL is done manually, a time- and effort- consuming process.

We are motivated in our study to extend the power of the framework to computational analysis. The difficulty in automating analysis of linguistic functions lies in both the fuzziness in the functional domain and a lack of relevant computational resources. The most significant lack of resource is a high-quality reference corpus crucial to statistical analysis and modeling. In the following sections, we discuss our initial efforts in constructing such a resource on a collaborative annotation platform and present the initial results from the corpus. The corpus is our first step in bridging the gap between the linguistic theory and application of such theory including automated analysis of language functions.

2. RELATED WORKS

Large-scale linguistic corpora have played an important role in natural language processing research and development of advanced machine learning algorithms. One important epistemological advantage of the corpus-based approach to linguistic study is that a high-quality annotated corpus provides a representative and systematic collection of empirical evidence for data mining. Corpus-based studies can uncover linguistic features that are inaccessible to intuition or cannot be deduced from a few small samples (Bednarek 2009). Over the past decades, the construction of prominent linguistic corpora to account for the syntactic

(Marcus 1993), semantic (Kingsbury et al. 2002) and discourse (Carlson and Okurowski 2002; Prasad et al. 2008) structures of linguistic information has deepened our understanding in each layer and made possible automated data-driven analysis based on them. Although the advantages of a functional-semantic orientation are apparent to text analysis, the complexity arising from annotation of multi-level functional-semantic information, such as that found in SFL, has led to a scarcity in large-scale, high-quality corpora annotated with such information (Honnibal and Curran 2007). While the possibility and suitability of SFL in its application to computational analysis have been duly discussed (Halliday and Webster 2006) and successfully applied in a number of NLP applications, particularly in Natural Language Generation (Teich 1999) a lack of high-quality SFL-based computational resources, especially a large-scale reference corpus, has impeded its applications in a wider range of problems.

A number of tools have been developed for annotating multi-layered functional structures, such as Genesys (Kumano et al. 1994), PALinkA (Orasan 2003) and UAM Corpus Tool (O'Donnell 2008). Despite addressing some of the difficulties in functional annotation, these tools still exhibit certain significant drawbacks such as: (1) inability to represent discontinuous and embedded units; (2) incompatibility with other annotation structures and formats; (3) lack of visualization of annotated structures; (4) overcomplicated interface; (5) nil collaboration among annotators; and (6) poor support for multi-language tagging. For example, the UAM CorpusTool, which represents the current state-of-the-art development of SFL annotation tools with a significant user base, is designed for single users working on a local computer with no embedded collaborative functionality, uses a propriety format for representing SFL annotations, and has a relatively steep learning curve[1].

Efforts have been made to circumvent the difficulties in manual annotation by attempting to convert the Penn Treebank to an SFL corpus (Honnibal and Curran 2007). The project has been partially successful in aligning basic functional components with syntactic structures in the Penn Treebank. It is argued that the partial success in converting the

[1] For example, one of our collaborators complained that although he finds the user manual for the tool helpful, he still has difficulty in understanding how to utilize the tool.

basic functional categories is due to the consistent annotation schemes of the Penn Treebank, and the SFL's remarkable agreement with other linguistic theories on the distinction of syntactic components, despite its emphasis on feature structures rather than syntactic representation. However, the work has been mostly concerned with the surface features of the SFL that are more or less syntactically oriented, while being unable to produce fine-grained functional-semantic categories that are crucial for any in-depth analysis of texts based on SFL. A high-quality functional corpus is still needed to fill this gap.

A number of linguistic resources annotated with shallow semantic roles have been produced over the years. Notable among them are the following three: FrameNet, VerbNet and Propbank.

The FrameNet database (Baker et al. 1998) is a semantic corpus annotated on the British National Corpus. The corpus annotates the frames of sentences using three components: lexicons, frames, and example sentences. Frames, or the context-sensitive conceptual structure, organized hierarchically, are composed of frame elements specific to a particular frame. Such annotations provide valuable context-specific knowledge and are useful for capturing certain semantic or syntactic patterns.

VerbNet (Schuler 2005) is a domain-independent verb lexicon with linkage to other lexical resources such as FrameNet and WordNet. It provides complete descriptions of verbs based on Levin's original classification (Levin et al. 1993), with substantial refinement. Each verb class in VerbNet is annotated with syntactic descriptions called syntactic frames, which define the surface realization of the predicate-argument structure for transitive, intransitive, prepositional phrases, etc., and thematic roles (e.g., Agent, Location, Theme) of its arguments. Semantic selectional restrictions (human, animate, organization, etc.) specify what thematic roles are allowed in the classes.

Propbank (Kingsbury and Palmer 2002) is another semantically-labeled resource. Annotated on one million words of the Wall Street Journal section of the Penn Treebank, it provides a detailed description of the predicate-argument structure of the annotated texts. The theoretical assumption underlying the annotations is fundamentally the same as that of the VerbNet: the semantics of sentences are reflected

in the syntactic frames associated with a verb of a particular verb class according to Levin's classification. The argument structures are labelled arg0, arg1, arg2, etc., based on the semantic role they play in a sentence and regardless of their syntactic positions. Thus in the sentences: *John broke the window*, and *The window broke*, although the window is the syntactic object in the first and subject in the second, it is given the same argument label. This allows us to capture the similarities in transitivity alternations in sentences that are syntactically different.

The annotation of such semantically oriented resources is an important contribution to the study of the complex phenomenon of language meanings. Each of them is grounded in a particular framework with certain assumptions, one more suited for certain applications than the others. However, to account for a fuller spectrum of the multifaceted nature of language meanings, multiple complementary resources are often linked and combined. With a focus on language functions (language use in context), the work on the proposed functional corpus provides an alternative view to the semantic and functional aspect of language that can be useful in problems and applications not directly targeted by those pre-existing resources, such as Critical Discourse Analysis and Automatic Text Generation.

An important justification for the need for an SFL-annotated corpus is that the systematic view to language in SFL, with a well-defined and unified taxonomy for the various facets of language, provides an ideal handle for the processing of linguistic information computationally. For example, in sentiment analysis (classification of texts based on their subjective sentiment), a task considered to be difficult to do computationally, the concepts of appraisal groups (Martin and White 2005) in SFL have helped computational linguists build the necessary semantic features that resulted in substantial improvement in sentiment analysis systems (Whitelaw et al. 2005).

Contextual information is essential to understanding the underlying functional- semantics in language. The standard representation of documents as bags of context- independent textual elements can also be augmented with the introduction of context-sensitive functional-semantic features which have been successfully applied to tasks such as stylistic text classicization (Argamon et al. 2007). Complementing a purely

event-based modeling on reality employed in frameworks such as FrameNet and VerbNet, the multifaceted view to language function models how human interactions are cohesively coded in language as interpersonal and textual metafunctions, which have proved to be essential in building natural language generation systems simulating human-like language patterns (Teich 1999).

3. CORPUS CONSTRUCTION

3.1 Text Selection

To leverage existing resources, the new corpus is annotated on the Penn Treebank with texts taken from the Wall Street Journal section. The same raw texts form a common basis of three well-established corpora: the Penn Treebank, the RST Discourse Treebank, and the Penn Discourse Treebank, making it possible for easy automatic alignment (establishing word-to-word correspondence) among the corpora. We align our functional-semantic features with each of these corpora to create a multilayered inter-linked information structure that can be used to explore the interactions and correlations of syntactic, discourse and functional information. At the grammatical layer, raw texts are annotated with part-of-speech and syntactic information in the Penn Treebank. On top of the same texts and grammatical annotations, two sub-layers of discourse relations describing the lexical ties and rhetorical structures have been annotated in the RST Treebank and the Penn Discourse Treebank. Aligned with the grammatical and discourse layers is the newly constructed functional layer in the proposed SFL corpus which annotates the three metafunctions of the same texts.

Table 1. The different layers of corpus information on the same textual base

Layer	**Sub-layer**	**Corpus**
Functional	Textual	The Proposed SFL Corpus
	Interpersonal	
	Experiential	
Discourse	Rhetorical	The RST Treebank
	Lexical	The Penn Discourse Treebank
Grammatical	Syntactic	The Penn Treebank
	Part-Of-Speech	
	Raw Text	

The PTB, RST-DT, PDTB are all annotated on top of the same raw text base. The PTB provides a solid syntactic ground on which other annotations are constructed. The PDTB annotates lower level theory-agnostic lexical connections while the RST-DT describes high-level discourse using a well-established discourse theory. Since they are so closely intertwined it is reasonable to study them in conjunction. It follows that if we are to build a new corpus that studies the correlation and interaction of functional and discourse information, it is preferable to build on this existing annotation base. It is possible to align our functional analysis with each of these corpora to create a multilayered information structure that can be used to explore the interactions and correlations within. Under a common database structure, links are built between these heterogeneous and traditionally incompatible datasets. In addition, as a theory about grammar and functions, SFL analysis can benefit from reference to syntactic and discourse properties of a text. Thus, although these resources appear to be heterogeneous, it is possible to utilize them using the same theoretical framework.

3.2 Corpus Details

Specific guidelines on the annotation task have been designed in accordance with the reference materials. The guidelines are stored in an online Google document. Project members with sufficient privileges are

allowed to create and modify the guidelines if such needs arise. The modifications made are visible to others members immediately.

The annotation was done in four successive layers, in which each of the following constituents is annotated:

Clausal: clausal boundaries, including boundaries of embedded clauses. The clause boundaries are aligned with the RST Treebank where clausal boundaries are also annotated, with fine-grained changes made to make it more suited for SFL's definitions of clauses.

Process: processes are the center of a clause, typically realized by a verbal group headed by the root verb of the clause. As described in (Halliday 1994; Martin et al. 2010), there are six common types of processes (*material, behavioral, mental, verbal, relational, existential*), subdivided into ten more refined types. Each of the process types is associated with a set of nuclear and non-nuclear participants.

Table 2. Major categories of process types and their category meanings.

Process Type	**Category Meaning**	**Examples（Process is underlined）**
Material	doing	
action	doing	*The car hit the tree.*
event	happening	*The snow melted.*
Behavioral	behaving	*He laughed.*
Mental	sensing	
perception	seeing	*I saw something.*
affection	feeling	*My son liked the toy car.*
cognition	thinking	*I think that's wrong.*
Verbal	saying	*He replied.*
Relational	being	
attribution	attributing	*The sky is blue.*
identification	identifying	*Obama is the President of the US.*
Existential	existing	*There is enough for everyone.*

Participant: participants are the central nominal groups of the clause typically realized by subject or objects of the clause. A summary of the process with its related participants is shown in Table 3.

Table 3. A summary of the process types and participants in the corpus, adapted from (Martin et al. 2010).

Process type	**Nuclear participants**	**Example**	**Additional non-nuclear participants**
material	Actor, Goal	*She made the coffee*	Initiator, Recipient, Client, Scope, Attribute
mental	Senser, Phenomenon	*She saw the car*	Inducer
relational: attributive	Carrier, Attribute	*Maggie was strong*	Attributor, Beneficiary
relational: identifying	Token, Value	*Maggie was our leader*	Assigner
behavioural	Behaver, (Target)	*she laughed*	Behaviour, Scope
verbal	Sayer, (Target)	*she replied*	Receiver, Verbiage
existential	Existent	*there was a beautiful princess*	

Circumstance: more-peripheral units related to time, place, manner, etc., typically realized by adverbial groups. There are in total nine broad types of circumstances: *Extent, Location, Manner, Cause, Contingency, Accompaniment, Role, Matter,* and *Angle,* each with its own subtypes. The *Extent* circumstance, for example, is subdivided into three subtypes: *duration, frequency,* and *distance*.

3.3 Annotation Infrastructure

The corpus is annotated using a web-based collaborative Tagger that we recently developed. The Tagger aims at providing a theory-neutral annotation framework for annotating heterogeneous (syntactic, semantic, functional, discourse) layers of linguistic information, multimodal data (e.g., images, sounds, videos) and metadata (e.g., user management, access control, time and geographical information).

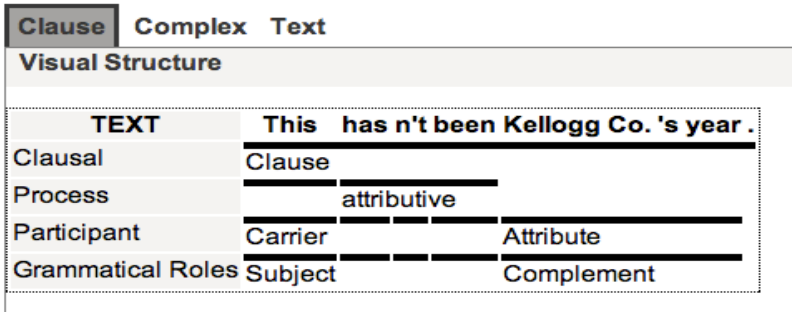

Figure 1. A structured view of a clause in the annotated corpus, taken from the web-based interface. In SFL, a process is typically realized by a verbal group, which can consist of lexical verbs (in this case *been*), finite verbs (*has*), polarity (*n't*) etc.

The Tagger is built on a generic, multifunctional database framework compatible with the Annotation Graph (Bird and Liberman 1999), an abstract annotation framework capable of representing a wide range of common linguistic signals (text, speech, image, video, multimodal interactions etc.), with properties particularly suited for collaborative annotation. Traditionally corpus annotation projects have been developed in isolation, leading to incompatibility in different annotation scheme and data storage formats. Later, realizing the importance of interoperability among different linguistic resources, efforts have been made, such as the Text Encoding Initiative (TEI) (Ide and Véronis 1995), to unify the way linguistic information is represented and stored. Since Bird and Liberman (2001), it has been generally accepted that generic linguistic annotations should be based on graphs. Recent state-of-the-art annotation frameworks, such as ATLAS (Bird et al. 2000) and LAF/GrAF (Ide and Suderman 2007), have used standoff XML formats that are interlinked and cross-referenced and thus impose a strict separation between linguistic data and graphs that encode annotations. The representation of annotated information as discussed in this paper follows this graph-based standard.

```
<?xml version="1.0" encoding="utf-8"?>
<graph xmlns="http://www.xces.org/ns/GrAF/1.0/">
    <graphHeader>
        <labelsDecl>
            <labelUsage label="sfl" occurs="3"/>
        </labelsDecl>
        <dependencies/>
        <annotationSpaces/>
    </graphHeader>
    <node xml:id="node1">
        <link targets="token1"/>
    </node>
    <region anchors="0 4" xml:id="token1"/>
    <a as="sfl" label="sfl" ref="node1" xml:id="a-0">
        <fs>
            <f name="theme">theme</f>
            <f name="transitivity">sayer</f>
            <f name="mood">subject</f>
        </fs>
    </a>
    <node xml:id="node2">
        <link targets="token2"/>
    </node>
    <region anchors="5 9" xml:id="token2"/>
    <a as="sfl" label="sfl" ref="node2" xml:id="a-1">
        <fs>
            <f name="theme">rheme</f>
            <f name="transitivity">process:verbal</f>
            <f name="mood">finite/predicator</f>
        </fs>
    </a>
    <node xml:id="node3">
        <link targets="token3"/>
    </node>
    <region anchors="10 19" xml:id="token3"/>
    <a as="sfl" label="sfl" ref="node3" xml:id="a-2">
        <fs>
            <f name="theme">rheme</f>
            <f name="transitivity">verbiage</f>
            <f name="mood">complement</f>
        </fs>
    </a>
    <node xml:id="node4">
        <link targets="token4"/>
    </node>
    <region anchors="20 21" xml:id="token4"/>
</graph>
```

Figure 2. LAF/GrAF XML serialization for SFG annotations on a simple clause. For a detailed description of the XML serialization, see (Ide and Suderman 2007).

In coding functional-semantic information, certain differences exist from coding formal structures. Many formal structures (e.g., syntactic trees in the Penn Treebank) can be economically represented as trees (a type of graph structure in which each node has a single parent), which presume a single hierarchical structure. Trees, however, are inadequate for representing functional-semantic structures due to their incompatibility with multiple parallel coding on the same raw text. Functionalist theories like SFL approach language from several multiple perspectives often leading to "redundancy" in representation. Such "redundancy" is often necessary to capture the multifaceted functional-semantic structures in language. Additionally, the presence of overlapping and discontinuous linguistic units in functional structures cannot be straightforwardly covered by a strict hierarchical tree. An adequate representation for functional structures should capture the richness and flexibility of the underlying theoretical assumptions in a uniform format.

We generalize the representation of functional-semantic information following the data model based on a directed graph. A graph of annotations *G*, as formally defined, is a set of vertices *V(G)* connected by a set of edges *E(G)*. Each vertex/edge can be labeled with one or more features. Each feature is a flexible mapping from a string to a value. The value in turn can be a string or a graph (the graph, called an *attribute-value graph*, is used to represent complex attribute-value relations). Each vertex in an annotation graph is an abstraction pointing to segments of linguistic signals.

This generic layered framework lends flexibility to alignment of noncontiguous words and other linguistic resources, useful for the nonconventional segmentation of functional components (such as the common anticipatory '*it*' as in "*It is a good thing that he stepped down as President.*") in SFL.

The Tagger features immediate annotation feedback through visualization, a process known to improve the quality and efficiency of annotation. For instance, when tagging at a particular layer (e.g., syntactic structure), information of the other layers (e.g., semantic properties) is immediately visible in a hierarchical structured format. This visualized information serves as additional references to the current

layer being annotated, especially when they are closely related in terms of function or meaning. When annotation errors (e.g., misalignment, mismatched labeling) are made they are immediately visible from the annotation interface for appropriate actions such as deletion or modification to be taken.

3.4 Quality Assurance

In the annotation process we follow a functional-semantic approach. Such an approach is characterized by its priority for cognitive judgment rather than linguistic criteria (Bhatia 1993; Kwan 2006). However, to strike a balance between flexibility and inter-annotator consistency and ensure the overall quality of the annotation, we adopt well-defined criteria for the functional categories and follow a number of quality-assurance procedures.

We adopt Halliday's seminal works (Halliday 1994; Halliday and Matthiessen 2004) on the theory to provide a standard reference due to the maturity and wide adoption of the works. Specific guidelines on the annotation task are designed in accordance with these reference materials.

Annotation quality and consistency are maintained by standard measures such as online documenting guidelines, training and tutorials, and multiple passes. In annotating functional-semantic features, we seek a balance by preserving reasonable alternative interpretations, while striving to reduce annotation errors. A logging and tracking mechanism is introduced that tracks all online activities in real time for supervisors to review annotation and provide real-time feedback to annotators for correction and improvement.

The tool uses a Wiki-like message board for discussions between annotators and public users, a process known to improve quality of collaborative knowledge construction (Kittur and Kraut 2010). Questions and feedback, along with a set of constantly updated guidelines, are recorded in a version-controlled database to be retrieved whenever needed and to guide new annotators and future annotations where similar scenarios arise. Each change made on the annotation tool is traceable, allowing for rollback at a later time (e.g., in case of a critical error). On

public-facing projects, it is an effective measure against potential vandalizers.

One major difficulty in ensuring the annotation quality of the proposed functional corpus lies in the inherent ambiguity in language functions. Even in a restricted context, there can be multiple interpretations of the same text. Unlike transformational grammars which study syntactic properties independent of context, functional theories such as SFL are grounded on the belief that language functions that a particular text serves can only be understood by taking into account all the related contextual factors, which are often culturally and socially dependent and subject to subjective interpretation. This leads to difficulties in disambiguating the meanings and functions of texts.

In annotating the functional corpus, the boundaries of some of the functional concepts are not always clear-cut. For example, apart from the three major functional types of process, material, mental and relational, there are three other types of processes that lie between the boundaries of any two of them: verbal, behavioral and existential. With such indeterminate boundaries, classification of the process types can often be difficult (see Section 4 for some examples). For the purpose of preserving alternative interpretations that also reflect the functional diversity of the structure, we choose to preserve multiple annotations of the same components. The annotations are ordered in terms of the perceived plausibility, resulting in primary annotations and secondary annotations that coexist.

User feedback is collected by sending out online surveys to continually improve the annotation interface, because user-friendliness of the annotation tool is found to be a key to maintaining accuracy, consistency and efficiency (Dukes et al. 2013).

A forking system is set up to maintain multiple copies of existing data. When collaborators wish to experiment with developing/modifying annotations of a project, they can safely (without affecting the original hosted data) fork the project to make changes, results from which can be merged back or discarded. In other words, an annotation project can be set to read-only while allowing it to be forked for experimentation and customization.

In annotating the corpus, small teams of three people are formed as an "annotation triangle", with one being the mentor whose role is to resolve any ambiguities or dissent between the other two, usually less experienced annotators. Annotators work in pairs. Texts are first annotated by one annotator from the pair, and subsequently reviewed by his partner. When an annotator disagrees with the other, the conflict is resolved by a third party – the mentor who makes the decision as to which interpretation to accept (sometimes both are maintained). A logging and tracking mechanism is introduced that tracks all online activities in real time, for supervisors to review annotation and provide real-time feedback to annotators for correction and improvement.

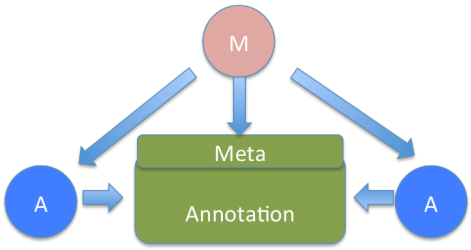

Figure 3. The Annotation Triangle where M (Mentor) and A (Annotators) work in Collaboration, where *Meta* refers to the settings related to project management and *Annotation* to the actual annotation work.

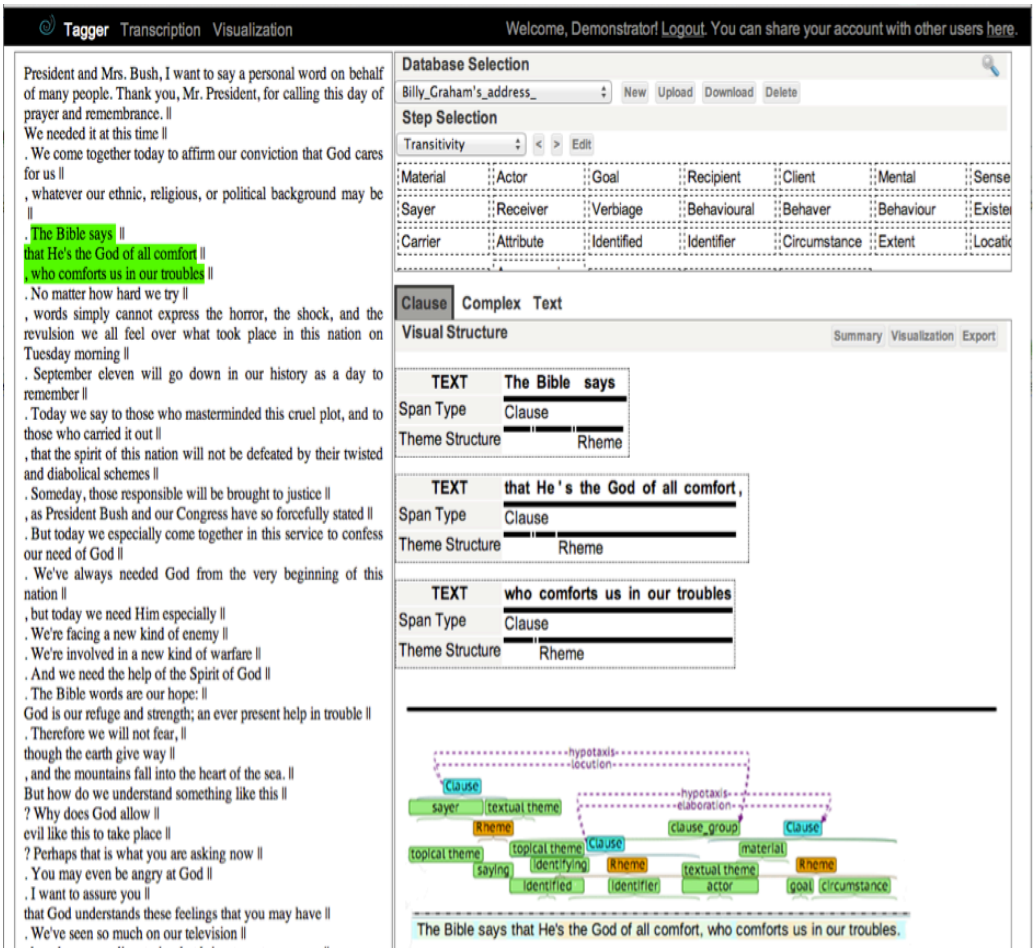

Figure 4. A view of the web-based collaborative tagger for annotating the functional and discourse structures of multilingual texts. The web-based interface is divided into three operation panels, namely, the text panel (left), annotation panel (top right) and visual structure panel (bottom right).

4. ANNOTATION STATISTICS

The construction of the functional corpus is an on-going project. The current corpus is constructed by a small team of annotators, all linguistic majors at graduate or undergraduate levels with formal training in the theoretical framework. After an initial three months of annotation we have constructed a small-scale corpus. In total we have annotated 81 documents from Wall Street Journal section of the Penn Treebank, with a total number of 43351 words, divided into 1621 sentences and 4620 clauses. The statistics of the top five types of annotated processes, participants, and circumstances are shown in Table 4.

In total, we have identified 912 verb types. The verb types are automatically identified by extracting the core verb from each verbal group in all annotated clauses in the corpus and then lemmatizing it using WordNet's Lemmatizer (Bird 2009). For example, in the clause *The movement is called a vibration*, the process, as realized by a verbal group, is *is called*, while the core verb in the verbal group is *called*, which is lemmatized to its base form *call*. In total, 218 word types have more than one process type (details of the number of each process type as represented by verb types is shown in Table 5). An uneven distribution of process types is noted in the statistics. The majority of the word types (714, or 76.6% of the total word types) are unambiguous, having only one process type, while the rest have one to six process types. Further work on the computational processing of functional structure will focus on disambiguating the word types with more than one functional meaning, taking into consideration the textual, grammatical and semantic contexts in which these different process types reside.

Table 4. Number of occurrences and percentage of each of the functional types.

Process Type	**Number**	**Percentage**
doing	1871	44.63%
happening	673	16.05%
verbal	585	13.96%
attributive	464	11.07%
identifying	216	5.15%
Participant Type	**Number**	**Percentage**
Goal	1608	23.85%
Actor	1300	19.28%
Verbiage	1153	17.10%
Sayer	517	7.67%
Attribute	469	6.96%
Circumstance Type	**Number**	**Percentage**
place	841	33.71%
quality	288	11.54%
degree	265	10.62%
guise	260	10.42%
comparison	125	5.01%

Table 5. Examples of the process call with four different process types.

Process Type	**Lexical Meaning**	**Example (processes are underlined)**
material: action	phoning somebody	*The president called him earlier tonight.*
relational: identification	identify; describe	*This movement is called a vibration.*
verbal	say loudly	*The butcher's son called out a greeting.*
mental: cognition	consider; regard	*This act can hardly be called generous.*

Table 6. Number of verb types and the number of process types that a verb type has.

Number of Process Types	1	2	3	4	5	6
Number of Verb Types	714	168	37	7	4	2

We calculate the inter-annotator agreement statistics on the three functional components: process types, participants and circumstances. We consider agreement to be cases where both the boundaries and types of functional labels are the same. The agreement ratio is 93.78% for process types, 87.47% for participants, and 86.13% for circumstances. The lower agreement in participants and circumstances is due to the fact that sometimes the boundaries of the structure that represent these functional components are not universally agreed upon. Although there is still room for improvement, the agreement is already high considering the fact that functional labels are often inherently more subjective than their lexical/syntactic counterparts.

5. CONCLUSION AND FUTURE WORK

In this paper, we discuss our work on constructing a functional corpus based on an influential theoretical framework. We present our initial attempts at building the corpus on a collaborative annotation platform. Although the scale of the functional corpus is still relatively small, its construction has made it possible to study basic functional properties computationally.

As an experiment, a prototypical classification system is built based on the annotated results for automatically classifying the functional processes of clauses using machine-learning algorithms such as Support Vector Machine (Tong and Koller 2002), results from which are to be presented in another paper. The potential use of the functional corpus is promising, with prospects of further developing into an important resource for carrying out fully automated functional analysis. The corpus and the experimental classifier will be further employed to build a large-scale functional corpus with substantially less effort. We plan to continue to expand the current corpus before releasing it to the community for researchers to further explore its potential application in a wide range of areas.

REFERENCES

Argamon, S., Whitelaw, C., Chase, P., Hota, S. R., Garg, N., and Levitan, S. 2007. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology* 58.6:802-822.

Baker, Collin F., Charles J. Fillmore, and John B. Lowe, 1998. The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, Université de Montréal, Montreal, pp.86-90.

Bednarek, M. 2009. Corpora and discourse: A three-pronged approach to analyzing linguistic data. In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus,* Sydney, Australia, pp.19-24.

Bhatia, V. K. 1993. *Analysing Genre: Language Use in Professional Settings*. London: Longman.

Bird, S. 2009. *Natural Language Processing with Python* (1st ed.). Beijing ; Cambridge [Mass.]: O'Reilly.

Bird, S., Day, D., Garofolo, J., Henderson, J., Laprun, C., and Liberman, M. 2000. ATLAS: A flexible and extensible architecture for linguistic annotation. *arXiv preprint cs/0007022*.

Bird, S., and Liberman, M. 1999. Annotation graphs as a framework for multidimensional linguistic data analysis. In *Towards Standards and Tools for Discourse Tagging–Proceedings of the Workshop,* University of Maryland, pp.23-60.

Bird, S., and Liberman, M. 2001. A formal framework for linguistic annotation. *Speech Communication* 33.1:23-60.

Carlson, L., Okurowski, M. E., and Marcu, D. 2002. *RST Discourse Treebank*. Linguistic Data Consortium, University of Pennsylvania.

Collins, M. 2003. Head-driven statistical models for natural language parsing. *Computational Linguistics* 29.4:589-637.

Dukes, K., Atwell, E., and Habash, N. 2013. Supervised collaboration for syntactic annotation of Quranic Arabic. *Language Resources and Evaluation* 47.1:33-62.

Gildea, D., and Jurafsky, D. 2002. Automatic Labeling of Semantic Roles. *Computational Linguistics* 28.3:245-288.

Halliday, M. A. K., and Matthiessen, C. 2004. *An Introduction to Functional Grammar.* London: Hoddar-Arnold.

Halliday, M. A. K. 1994. *An Introduction to Functional Grammar* (2nd ed.). London: Edward Arnold.

Halliday, M. A. K. 2005. *Computational and Quantitative Studies*, Vol. 6. London: Bloomsbury Publishing.

Hernault, H., Prendinger, H., DuVerle, D. A., and Ishizuka, M. 2010. HILDA: A discourse parser using support vector machine classification. *Dialogue and*

Discourse 1.3:1-33.

Honnibal, M., and Curran, J. R. , 2007. Creating a systemic functional grammar corpus from the Penn treebank. In *Proceedings of the Workshop on Deep Linguistic Processing*, Prague, Czech Republic, pp.89-96.

Ide, N. M., and Véronis, J. 1995. *Text Encoding Initiative: Background and Contexts*. Dordrecht; Boston: Kluwer Academic Publishers.

Ide, N., and Suderman, K. , 2007. GrAF: A graph-based format for linguistic annotations. In *Proceedings of the Linguistic Annotation Workshop*, Prague, Czech Republic, pp.1-8.

Kingsbury, P., and Palmer, M., 2002. From treebank to propbank. In *Proceedings of the Third International Conference on Language, Resources and Evaluation, The University of Las Palmas de Gran Canaria, Las Palmas, Spain,* pp.1989-1993.

Kingsbury, P., Palmer, M., and Marcus, M, 2002. Adding semantic annotation to the penn treebank. In *Proceedings of the Human Language Technology Conference*, San Diego, California, USA, pp.252-256.

Kittur, A., and Kraut, R. E., 2010. Beyond Wikipedia: coordination and conflict in online production groups. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, Savannah, Georgia, USA, pp.215-224.

Klein, D., and Manning, C. D. C., 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics ACL 03*, *1*(July), Sapporo, Japan, pp.423-430.

Kumano, T., Tokunga, T., Inui, K., and Tanaka, H. , 1994. GENESYS: An integrated environment for developing systemic functional grammars. In *Proceedings of the International Workshop on Sharable Natural Language Resources, Ikoma, Nara, Japan,* pp.78-85.

Kwan, B. S. 2006. The schematic structure of literature reviews in doctoral theses of applied linguistics. *English for Specific Purposes* 25.1:30-55.

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

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistic*s 19.2:313-330.

Martin, J. R., Matthiessen, C. M., and Painter, C. 2010. *Deploying Functional Grammar*. Shanghai: Commercial Press.

Martin, J. R., and White, P. R. 2005. *The Language of Evaluation*. Basingstoke and New York: Palgrave Macmillan.

O'Donnell, M., 2008. Demonstration of the UAM CorpusTool for text and image annotation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session*, Columbus, Ohio, USA, pp.13-16.

Orasan, C., 2003. PALinkA: A highly customisable tool for discourse annotation. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialog, Sapporo, Japan*, pp.39-43.

Pradhan, S. S., Ward, W., Hacioglu, K., Martin, J. H., and Jurafsky, D., 2004. Shallow

semantic parsing using support vector machines. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004,* Boulder, Colorado, USA, pp.233-240.

Schuler, K. K. 2005. *Verbnet: A Broad-coverage, Comprehensive Verb Lexicon*. University of Pennsylvania, Philadelphia, PA, USA.

Teich, E. 1999. *Systemic Functional Grammar and Natural Language Generation*. London: Bloomsbury Publishing.

Tong, S., and Koller, D. 2002. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research* 2:45-66.

Whitelaw, C., Garg, N., and Argamon, S., 2005. Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management,* Bremen, Germany, pp.625-631.

[Received 17 March 2014; revised 08 July 2014; accepted 16 October 2014]

Hengbin Yan
Faculty of English Language and Culture
Guangdong University of Foreign Languages
Guangzhou, Guangdong, People's Republic of China
yhb@gdufs.edu.cn

功能語料庫的一體化構建方法

嚴恒斌[1]、Jonathan Webster[2]
廣東外語外貿大學[1]
香港城市大學[2]

本文論述作者基於系統功能語法框架，構建一個全新語料庫的經驗。我們從Penn Treebank語料庫中選取部份文本，通過一個基於網絡且有著多項高級特性的協作性平台對文本進行標註。我們首先討論我們項目的背景和目的，然後提出我們針對協作性標註過程中所遇到的一些問題和挑戰的解決方法。我們初步構建的語料庫有著較為精確的高質量標註，可對現有的基於語義標註的語料庫資源作有益的補充，同時也為進一步開發相關的大型功能語言學資源乃至語言功能自動分析系統的構建打下基礎。

關鍵字：語料庫標註、語言功能、協作性標註、功能語義