A Corpus-Based Approach to Linguistic Function

Hengbin Yan and Jonathan Webster
The Halliday Centre for Intelligent Applications of Language Studies
Department of Chinese, Translation and Linguistics
City University of Hong Kong
{hbyan2, ctjjw}@cityu.edu.hk

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 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 aid in the development of a larger-scale resource crucial for automated systems for analysis of linguistic function.

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 is becoming increasing influential in recent years. SFL provides an ideal handle to 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

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 refer-

27th Pacific Asia Conference on Language, Information, and Computation pages 215－221

ence corpus, has impeded its applications in wider range of problems.

A number of tools have been developed for annotating multi-layered functional structures, such as Gensys (Kumano et al., 1994), PALinkA (Orasan, 2003) and UAM CorpusTool (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) over-complicated interface; (5) nil collaboration among annotators; and (6) poor support for multi-language tagging.

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 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, 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 detailed description of the predicate-argument structure of the annotated texts. The theoretical assumption underlying the annotations are 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 structure 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 important contributions to the study of the complex phenomenon of language meanings. Each of them is grounded on 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.

3 Corpus Construction

3.1 Corpus Texts

To leverage existing resources, the new corpus is annotated on the Penn Treebank (Marcus, 1993) 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 (Carlson and Okurowski, 2002), and the Penn Discourse Treebank (Prasad et al., 2008), 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 interlinked information structure that can be used to explore the interactions and correlations of syntactic, discourse and functional information.

3.2 Annotation Infrastructure

The corpus is annotated using a web-based collaborative Tagger that we recently developed (see Figure. 2). 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).

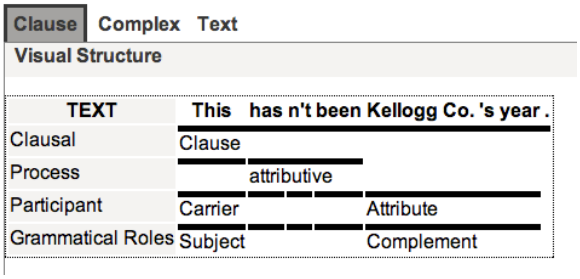

Figure 1: A structured view of a clause in the annotated corpus, taken from the web-based interface.

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. This generic layered framework lends flexibility to alignment of noncontiguous words and other linguistic resource, 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.

3.3 Quality assurance

Annotation quality and consistency are maintained by standard measures such as online documenting guidelines, trainings 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).

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 is grounded on the belief that language functions that a particular text serves can only be seen 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 language meanings and

functions. 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.

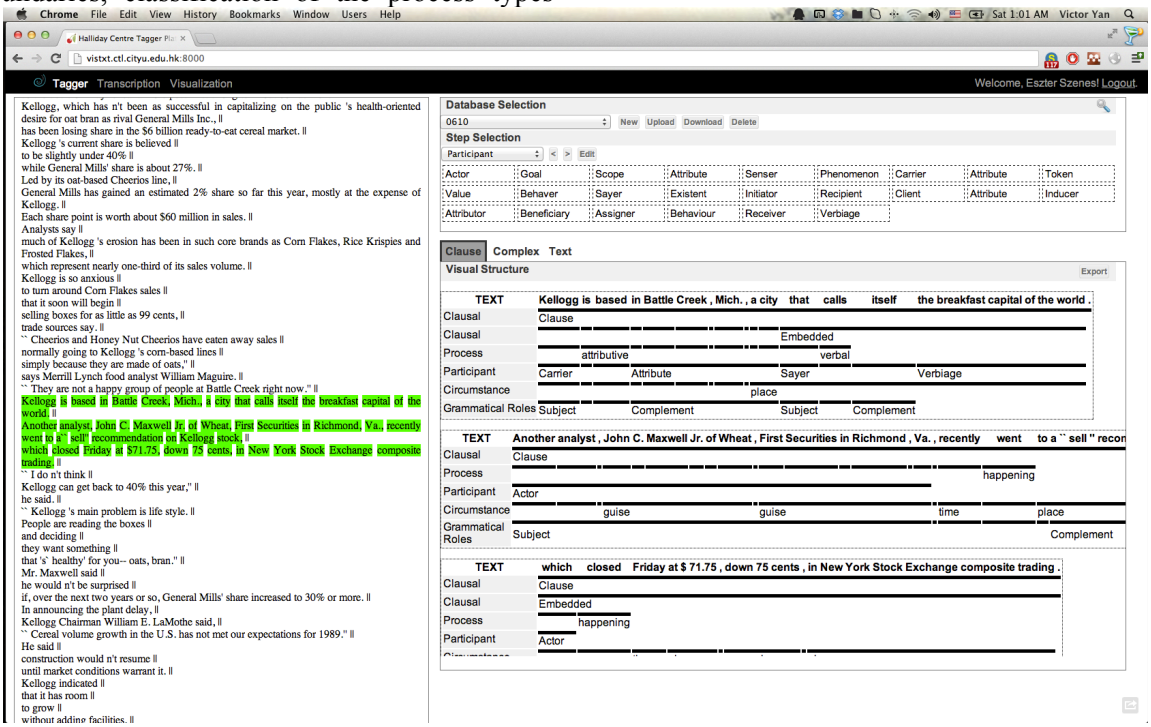

Figure 2: A view of the web-based collaborative tagger for annotating the functional and discourse structures of multilingual texts. The web-based interface is divided into three operation panels, namely, the text panel (left), annotation panel (top right) and visual structure panel (bottom right).

3.4 Corpus Details

We adopt Halliday's seminal works (Halliday, 1994) on the theory to provide 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.

In SFL-based analysis, three strata of meaning (called metafunctions) operate in parallel: the ideational, interpersonal, and textual metafunctions. As the other two layers are more syntactically oriented and convertible from syntactically parsed trees, we focus our annotation on the ideational metafunction, and more specifically a major subcategory, the experiential metafunction, whose categorization is largely functionally oriented and less lexically/syntactically dependent.

The experiential metafunction, as its name suggests, has to do with functions relating to world experience. Linguistically, it involves a configuration of processes and participants involved (such as Actor, Goal), and the accompanying circumstances (such as time, place, manner). Such configuration allows one to look beyond the sentence surface to probe into the semantic aspect of the text.

The annotation was done in three 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), 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. A summary of the process with its related participants is given in Table 1.

Participant: participants are the central nominal groups of the clause typically realized by subject or objects of the clause.

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*.

Process type	Nuclear participants	Non-nuclear participants
material: action event	Actor, Goal	Initiator, Recipient, Client, Scope
mental: perception affection cogition	Senser, Phenomenon	Inducer
Relational: attributive	Carrier, Attribute	Attributor, Beneficiary
identifying	Token, Value	Assigner
behavioural	Behaver, Target	Behaviour, Scope
verbal	Sayer, Target	Receiver, Verbiage
existential	Existent	

Table 1: A summary of the process types and participants in the corpus

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 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 2.

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 2: Number of occurrences and percentage of each of the functional types

In total, we have identified 912 verb types. The verb types are identified by extracting the core verb from each verbal group and then lemmatized using WordNet's lemmatizer (Bird et al., 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 are shown in Table 4).

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 3: Examples of the process *call* with four different process types

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

Table 4: Number of verb types and the number of process types that a verb type has.

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 definitely 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 & 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

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1* (pp. 86–90).

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'reilly.

Bird, S., & Liberman, M. (1999). Annotation graphs as a framework for multidimensional linguistic data analysis. In T*owards Standards and Tools for Discourse Tagging–Proceedings of the Workshop*.

Carlson, L., & Okurowski, M. (2002). RST discourse treebank.

Collins, M. (2003). Head-driven statistical models for natural language parsing. *Computational linguistics*.

Gildea, D., & Jurafsky, D. (2002). Automatic Labeling of Semantic Roles. *Computational Linguistics*, 28(3), 245–288.

Halliday, M., & Webster, J. (2006). *Computational and quantitative studies*.

Halliday, Michael A., & Matthiessen, C. M. (2004). An introduction to functional grammar.

Halliday, Michael AK. (1994). An Introduction to Functional Grammar. London: Edward Arnold.

Hernault, H., Prendinger, H., DuVerle, D. A., & Ishizuka, M. (2010). HILDA: A Discourse Parser Using Support Vector Machine Classification. *Dialogue & Discourse*, 1(3), 1–33.

Honnibal, M., & Curran, J. J. R. (2007). Creating a systemic functional grammar corpus from the Penn treebank. *Proceedings of the Workshop on Deep Linguistic Processing - DeepLP '07*, (June 2005), 89.

Kingsbury, P, Palmer, M., & Marcus, M. (2002). Adding semantic annotation to the penn treebank. *Proceedings of the Human Language Technology Conference*.

Kingsbury, Paul, & Palmer, M. (2002). From TreeBank to PropBank. In *LREC*.

Kittur, A., & 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* (pp. 215–224).

Klein, D., & Manning, C. D. C. (2003). Accurate unlexicalized parsing. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics ACL 03*, 1(July), 423–430.

Kumano, T., Tokunga, T., Inui, K., & Tanaka, H. (1994). GENESYS: An integrated environment for developing systemic functional grammars. In *Proceedings of the International Workshop on Sharable Natural Language Resources* (pp. 78–85).

Levin, B. (1993). English verb classes and alternations: A preliminary investigation (Vol. 348). University of Chicago press Chicago.

Marcus, M. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*.

O'Donnell, M. (2008). Demonstration of the UAM CorpusTool for text and image annotation. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies Demo Session - HLT '08*, (June), 13–16.

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

Pradhan, S., Ward, W., & Hacioglu, K. (2004). Shallow semantic parsing using support vector machines. *Proceedings of HLT/NAACL*, 233.

Prasad, R., Dinesh, N., & Lee, A. (2008). The penn discourse treebank 2.0. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, 2961–2968.

Schuler, K. K. (2005). VerbNet: A broad-coverage, comprehensive verb lexicon.

Teich, E. (1999). Systemic functional grammar in natural language generation: Linguistic description and computational representation.

Tong, S., & Koller, D. (2002). Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2, 45–66.