The Halliday Centre Tagger: An Online Platform for Semi-automatic Text Annotation and Analysis

Billy T.M. Wong,[1] Ian C. Chow,[2] Jonathan J. Webster,[3] Hengbin Yan[4]
[1,2]Department of Translation, The Chinese University of Hong Kong
Shatin, NT, Hong Kong SAR
[3,4]The Halliday Centre for Intelligent Applications of Language Studies, City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong SAR
[1]billywtm@gmail.com, [2]ianchow@arts.cuhk.edu.hk, [3]ctjjw@cityu.edu.hk, [4]hbyan2@cityu.edu.hk

Abstract

This paper reports the latest development of The Halliday Centre Tagger (the Tagger), an online platform provided with semi-automatic features to facilitate text annotation and analysis. The Tagger features a web-based architecture with all functionalities and file storage space provided online, and a theory-neutral design where users can define their own labels for annotating various kinds of linguistic information. The Tagger is currently optimized for text annotation of Systemic Functional Grammar (SFG), providing by default a pre-defined set of SFG grammatical features, and the function of automatic identification of process types for English verbs. Apart from annotation, the Tagger also offers the features of visualization and summarization to aid text analysis. The visualization feature combines and illustrates multi-dimensional layers of annotation in a unified way of presentation, while the summarization feature categorizes annotated entries according to different SFG systems, i.e., transitivity, theme, logical-semantic relations, etc. Such features help users identify grammatical patterns in an annotated text.

Keywords: The Halliday Centre Tagger, Systemic Functional Grammar, Corpus Annotation

1. Introduction

Annotation is a process to enrich a text with linguistic information which is implicitly present, that becomes a resource reusable for tasks such as language study and development of natural language processing (NLP) technology. Many available corpora, such as the Penn Treebank,[1] British National Corpus[2] and American National Corpus,[3] are annotated with basic linguistic information like part-of-speech, name entities and syntactic structures, with the aid of automatic taggers.

Our current work lies in the development of a corpus annotated according to Systemic Functional Grammar (SFG) (Yan & Webster, 2013). SFG describes the realization of meaning in language through a paradigmatic set of functional-semantic choices in the functional-semantic aspect. The annotation of SFG involves a multi-dimensional analysis of text based on three meta-functions, i.e., ideational, interpersonal and textual, each representing a layer of meaning with a set of options for annotators to pick. Although there are a few tools for SFG-annotation, such as Systemic Coder (O'Donnell, 1995), SysFan (Wu, 2000), LBIS Coder (Sugimoto et al., 2005) and UAM Corpustool (O'Donnell, 2008), they are limited by virtue of being standalone offline applications, offering no support for collaboration, and providing no features to assist manual annotation, thus highlighting the need for a better tool for the task.

The Halliday Centre Tagger (the Tagger)[4] (Chan et al., 2012; Yan & Webster, 2012; Wong et al., 2013) is an online platform provided with semi-automatic features to facilitate SFG-based text annotation and analysis. Annotation is performed online with support for collaboration. In order to reduce annotation efforts, some grammatical options are semi-automatically identified. The annotated entries in a text can be visualized and summarized, to help users intuitively identify and locate the occurrence of patterning.

2. The Halliday Centre Tagger

The Tagger is featured for its web-based architecture. A user simply needs to upload a text to the Tagger and then use available functionalities to annotate. All texts and annotation are stored online in the user's account and are web-accessible. Collaborative annotation is supported on the same text by multiple users either synchronously or asynchronously. Although the Tagger is optimized for SFG-annotation, it remains theory-neutral. Apart from a pre-defined set of SFG grammatical features provided by default, users can also define their own labels for annotating other types of linguistic information (Figure 1).

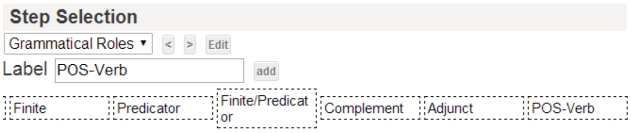
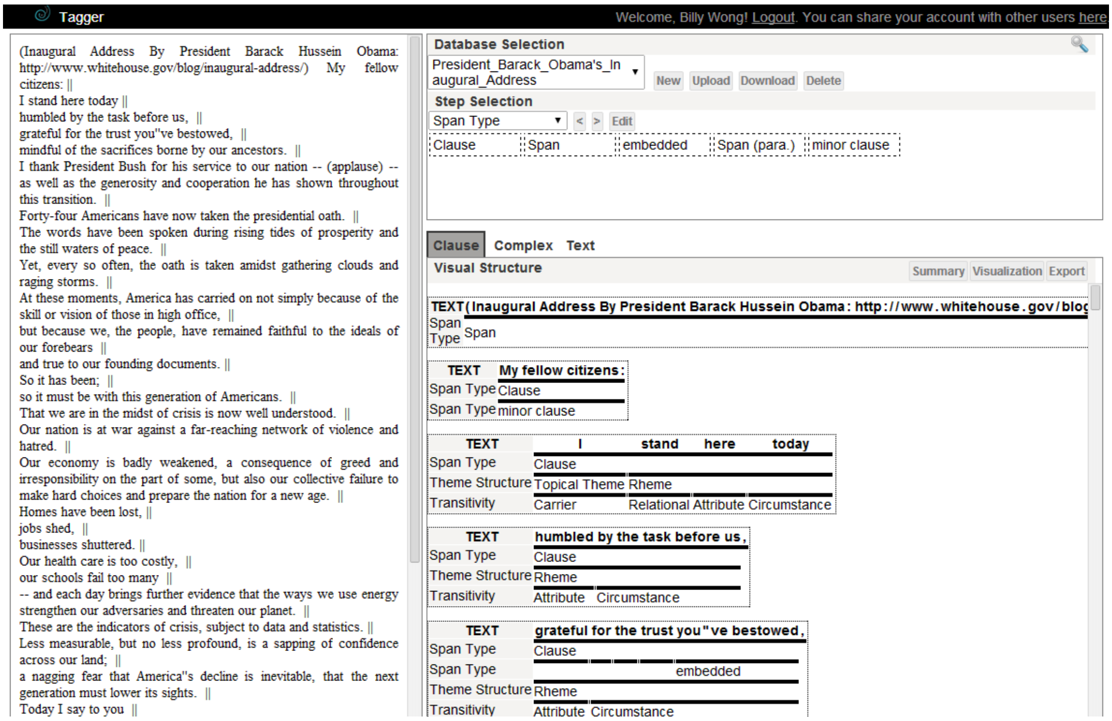

Step Selection
Grammatical Roles
Label POS-Verb

Figure 1: User-defined labels for annotation

The Tagger is also designed for supporting multi-dimentional annotation with different kinds of lingustic information annotated on the same text. They are represented as layers below a text span, each showing a type of linguistic information (Figure 2).

[1] http://catalog.ldc.upenn.edu/LDC99T42
[2] http://www.natcorp.ox.ac.uk
[3] http://www.americannationalcorpus.org
[4] http://hallidaycentre.cityu.edu.hk/06_hctagger.html



Welcome, Billy Wong! Logout. You can share your account with other users here.

President_Barack_Obama's_Inaugural_Address

(Inaugural Address By President Barack Hussein Obama: http://www.whitehouse.gov/blog/inaugural-address/) My fellow citizens: ||
I stand here today ||
humbled by the task before us, ||
grateful for the trust you''ve bestowed, ||
mindful of the sacrifices borne by our ancestors. ||
I thank President Bush for his service to our nation -- (applause) -- as well as the generosity and cooperation he has shown throughout this transition. ||
Forty-four Americans have now taken the presidential oath. ||
The words have been spoken during rising tides of prosperity and the still waters of peace. ||
Yet, every so often, the oath is taken amidst gathering clouds and raging storms. ||
At these moments, America has carried on not simply because of the skill or vision of those in high office, ||
but because we, the people, have remained faithful to the ideals of our forebears ||
and true to our founding documents. ||
So it has been; ||
so it must be with this generation of Americans. ||
That we are in the midst of crisis is now well understood. ||
Our nation is at war against a far-reaching network of violence and hatred. ||
Our economy is badly weakened, a consequence of greed and irresponsibility on the part of some, but also our collective failure to make hard choices and prepare the nation for a new age. ||
Homes have been lost, ||
jobs shed, ||
businesses shuttered. ||
Our health care is too costly, ||
our schools fail too many ||
-- and each day brings further evidence that the ways we use energy strengthen our adversaries and threaten our planet. ||
These are the indicators of crisis, subject to data and statistics. ||
Less measurable, but no less profound, is a sapping of confidence across our land; ||
a nagging fear that America''s decline is inevitable, that the next generation must lower its sights. ||
Today I say to you ||

TEXT	(Inaugural Address By President Barack Hussein Obama: http://www.whitehouse.gov/blog
Span Type	Span

TEXT	My fellow citizens:
Span Type	Clause
Span Type	minor clause

TEXT	I	stand	here	today
Span Type	Clause			
Theme Structure	Topical Theme	Rheme		
Transitivity	Carrier	Relational	Attribute	Circumstance

TEXT	humbled by the task before us,
Span Type	Clause
Theme Structure	Rheme
Transitivity	Attribute Circumstance

TEXT	grateful for the trust you"ve bestowed,
Span Type	Clause
Span Type	embedded
Theme Structure	Rheme
Transitivity	Attribute Circumstance

Figure 2: The annotation interface of the Halliday Centre Tagger

3. Semi-automatic Annotation

Manual SFG annotation requires considerable human effort as it involves multi-dimensional text analysis. Current progress in development of automatic SFG parsing remains rudimentary. Nevertheless, based on our previous work, we have developed and integrated into the Tagger some semi-automated features intended to improve the productivity of annotation.

SFG annotation involves text analysis in terms of the three meta-functions: ideational, interpersonal and textual. Each has its own set of grammatical and semantic categories denoting different underlying functions from different perspectives.

We focus on semi-automatic analysis of ideational meta-function which consists of logical and experiential meanings. The experiential focuses on construing the flux of experience and is structurally realized by the system of transitivity. Taking the clause as the most basic lexico-grammatical unit, clause constituents include a process, possibly one or more participants and circumstances. Process is typically a verb, defining the type of experiential meaning and governing the semantic roles of participants.

3.1 Process Type Identification

We provide the Tagger with the feature of semi-automatic identification of process types for English verbs. This is based on a lexicographical database of SFG process types developed by Chow (2008), in which the process type of each English verb-sense in WordNet[5] is identified via utilizing various interoperable lexical and ontological resources including GUM,[6] FrameNet[7] and SUMO[8] (Chow & Webster, 2007, 2008). For the English verbs in WordNet, over 95% of verb-senses are identified with corresponding SFG process types.

The identification of process type for an English verb begins with word sense disambiguation (WSD), in order to first identify the possible sense(s) of the verb based on its occurrence context. We employ a WordNet-based WSD system WordNet::SenseRelate:: AllWords (SR-AW) (Pedersen & Kolhatka, 2009) for this purpose, which offers a satisfying performance in determining word sense.[9] The process type of each possible word sense is then identified with the use of the lexicographical database.

The feature of process type identification is provided on the Tagger by eliminating (i.e., greying out) the improbable options of process type when annotating English verbs. Upon user's selection of a correct process type, the improbable options of semantic roles (participants) in the annotating clause which are dependent on the choice of process type are then greyed out. Users only need to pick out the correct ones from a reduced set of options, thus substantially reducing manual effort in doing annotation, especially with respect to semantic roles whose possible options are greater than twenty.

An experiment was performed to verify the effectiveness of this semi-automatic features. In our current annotation

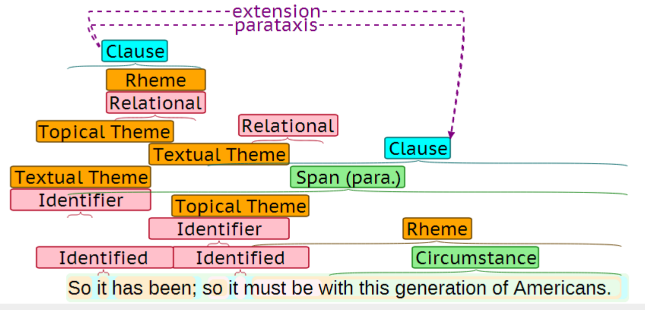
[5] http://wordnet.princeton.edu

[6] http://glotta.ntua.gr/StateoftheArt/Ontologies/newUM.html

[7] https://framenet.icsi.berkeley.edu/fndrupal

[8] http://www.ontologyportal.org

[9] According to Pedersen and Kolhatka (2009), SR-AW yields the F-measure results of 61% in SemCor, 59% in SSENSEVAL-2 and 54% in SENSEVAL-3.



So it has been; so it must be with this generation of Americans.

Figure 3: The visualization feature

project, a set of annotated speeches of 2,426 words were selected as test data, which contain totally 238 clauses. For process type identification, 167 out of 231 process types were correctly identified, i.e., an accuracy rate of 72%. The major errors come from the WSD system, i.e., parts-of-speech are incorrectly identified or phrasal verbs cannot be recognized.

4. Features for Text Analysis

Apart from annotation, the Tagger also provides the features of visualization and summarization to aid text analysis based on the annotation. Through reorganizing the presentation of annotated entries, these features help users to identify grammatical patterns in the annotated texts.

4.1 Visualization

The visualization feature combines and illustrates the multi-dimensional annotation in a unified way of presentation. Based on the Brat rapid annotation tool (Stenetorp et al., 2012), the multi-dimensional meta-functions in a text are now visualized in a color-coded, interactive and presentable manner (Figure 3). It shows the distribution of different SFG annotated entries in a text that helps one to intuitively identify and locate the occurrence of patterning.

4.2 Summarization

The summarization feature categorizes annotated entries according to different SFG systems, i.e., transitivity, theme, logical-semantic relations, etc. The use of various constituents in a text is systematically summarized, with statistics showing their occurrence frequency (Figure 4). Users can opt for an overall or focused summary depending on their interest in the use of all or particular constituents. The summary is provided in a table format popularly used in the study of SFG, saving users conceivable time in preparing this kind of material on their own.

Both features are fully automatic requiring no human intervention. Users only need to work on the annotation. The visualized and summarized results can then be generated by the system.

5. Summary

The latest development of the Halliday Centre Tagger has increased the automatic capability of the SFG annotation tool. The web-based architecture provides users' convenience in accessing the Tagger and managing annotation tasks online, and supports collaboration on annotation with other users. The theory-neutral design allows users to flexibly define their own labels for annotation of any kind of linguistic information. The feature of semi-automatic process type identification can raise users' productivity. The features of visualization and summarization aid users to carry out text analysis, in addition to merely doing annotation work. The Tagger is currently employed in our project of developing an SFG-annotated corpus for corpus-based study and development of related NLP technology for SFG, in which the present progress remains rudimentary. The ongoing progress and some intermediate statistics will be presented.

6. Acknowledgements

The research described in this paper was supported by City University of Hong Kong through the SRG grants 7002802 and 7004163.

7. References

Chan, C.L., Yan, H., Lee, S.Y., Webster, J. and Wong, H.K. (2012). A database design for complex linguistic data in collaborative Web application. In *Proceedings of the Second International Conference on Digital Information and Communication Technology and it's Applications (DICTAP)*. Bangkok, Thailand, pp. 159--165.

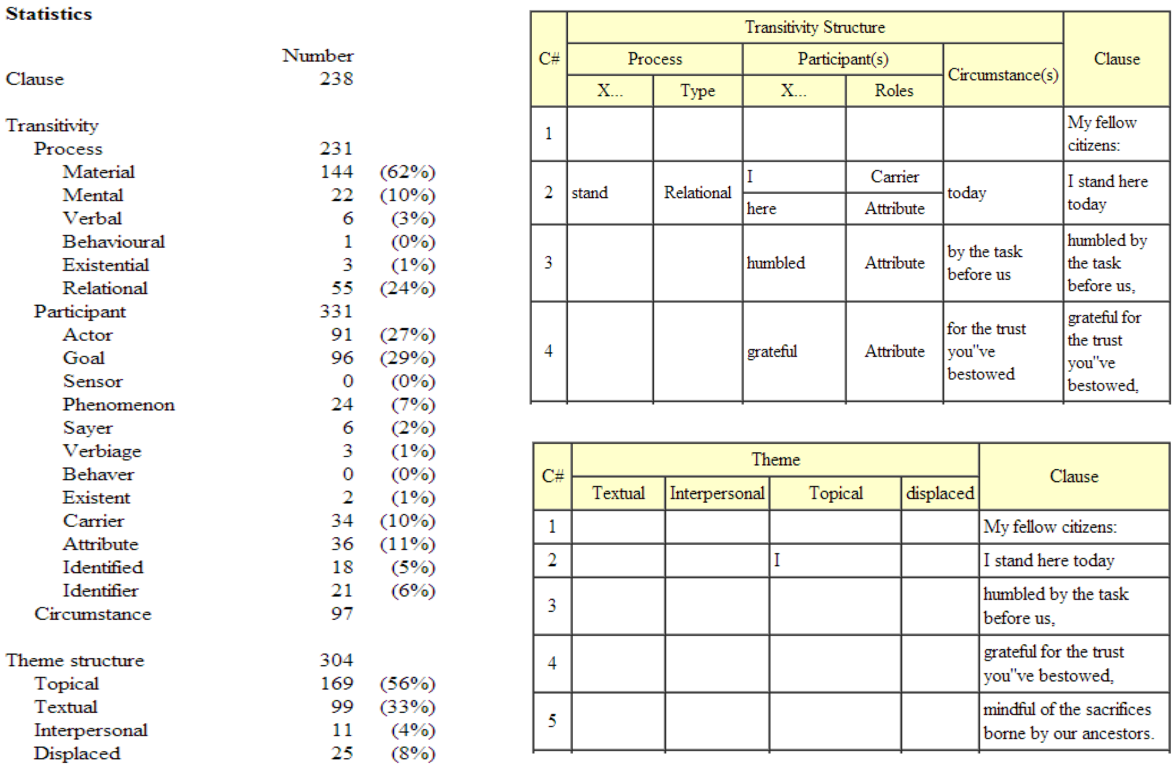
Chow, I.C. (2008). *Constructing a Linguistic Resource of Verbs: An Ontology Engineering Approach*. PhD dissertation. City University of Hong Kong: Hong Kong.

Chow, I.C. and Webster, J.J. (2007). Integration of linguistic resources for verb classification: FrameNet frame, WordNet verb and SUMO. *Lecture Notes of Computer Science LNCS*, Vol. 4394, pp. 1--11.

Chow, I.C. and Webster, J.J. (2008). Supervised clustering of the WordNet verb hierarchy for Systemic Functional Process type identification. In *Proceedings of International Conference on Global Interoperability for Language Resources (ICGL 2008)*. Hong Kong, pp. 51--58.

O'Donnell, M. (1995). From corpus to codings: Semi-automating the acquisition of linguistic features. In *Proceedings of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*. Stanford University, California.

O'Donnell, M. (2008). The UAM CorpusTool: Software for corpus annotation and exploration. In Bretones Callejas, Carmen M. et al. (eds) *Applied Linguistics Now: Understanding Language and Mind* / La Lingüística Aplicada Hoy: Comprendiendo el Lenguaje y la Mente. Almería: Universidad de Almería, pp. 1433--1447.



Statistics

	Number	
Clause	238	
Transitivity		
Process	231	
Material	144	(62%)
Mental	22	(10%)
Verbal	6	(3%)
Behavioural	1	(0%)
Existential	3	(1%)
Relational	55	(24%)
Participant	331	
Actor	91	(27%)
Goal	96	(29%)
Sensor	0	(0%)
Phenomenon	24	(7%)
Sayer	6	(2%)
Verbiage	3	(1%)
Behaver	0	(0%)
Existent	2	(1%)
Carrier	34	(10%)
Attribute	36	(11%)
Identified	18	(5%)
Identifier	21	(6%)
Circumstance	97	
Theme structure	304	
Topical	169	(56%)
Textual	99	(33%)
Interpersonal	11	(4%)
Displaced	25	(8%)

C#	Transitivity Structure					Clause
	Process		Participant(s)		Circumstance(s)	
	X...	Type	X...	Roles		
1						My fellow citizens:
2	stand	Relational	I here	Carrier Attribute	today	I stand here today
3			humbled	Attribute	by the task before us	humbled by the task before us,
4			grateful	Attribute	for the trust you''ve bestowed	grateful for the trust you''ve bestowed,

C#	Theme				Clause
	Textual	Interpersonal	Topical	displaced	
1					My fellow citizens:
2			I		I stand here today
3					humbled by the task before us,
4					grateful for the trust you''ve bestowed,
5					mindful of the sacrifices borne by our ancestors.

Figure 4: The Summarization feature

Pedersen, T. and Kolhatkar, V. (2009). WordNet:: SenseRelate::AllWords—A broad coverage word sense tagger that maximimizes semantic relatedness. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies Conference (NAACL-HLT 2009)*. Boulder, CO, pp. 17--20.

Sugimoto, T., Ito, N., Iwashita, S. and Sugeno, M. (2005). A computational framework for text processing based on systemic functional linguistics. In *Proceedings of the 1st Computational Systemic Functional Grammar Conference*. University of Sydney, pp. 2--11.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S. and Tsujii, J. (2012). Brat: a Web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Avignon, France, pp. 102--107.

Wong, B. T.M., Pun, C. F.K. and Webster, J.J. (2013). The Halliday Centre Tagger for corpus-based study of Systemic Functional Linguistics. In *Proceedings of the 7th Annual International Free Linguistics Conference*, Hong Kong.

Wu, C. (2000) *Modelling Linguistic Resources: A Systemic-functional Approach*. PhD thesis, Department of Linguistics, Macquarie University, Sydney, Australia.

Yan, H., and Webster, J. (2012). Collaborative annotation and visualization of functional and discourse structures. In *Proceedings of the 24th Conference on Computational Linguistics and Speech Processing (ROCLING)*.Taiwan, pp. 366--374.

Yan, H., and Webster, J. (2013). A corpus-based approach to linguistic function. In *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27)*. Taiwan, pp. 215--221.