Engineering an Aligned Gold-Standard Corpus of Human to Machine Oriented Controlled Natural Language

Conference Paper · December 2018

DOI: 10.1109/WI.2018.00-58

3 authors:

Hazem Safwat
National University of Ireland, Galway
6 PUBLICATIONS **19** CITATIONS
SEE PROFILE

Brian Davis
National University of Ireland, Galway
68 PUBLICATIONS **625** CITATIONS
SEE PROFILE

Manel Zarrouk
National University of Ireland, Galway
37 PUBLICATIONS **146** CITATIONS
SEE PROFILE

2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)

Engineering an aligned gold-standard corpus of human to machine oriented Controlled Natural Language

Hazem Safwat
Insight Centre for Data Analytics
National University of Ireland, Galway
hazem.abdelaal@insight-centre.org

Brian Davis
Insight Centre for Data Analytics
National University of Ireland, Galway
brian.davis@insight-centre.org

Manel Zarrouk
Insight Centre for Data Analytics
National University of Ireland, Galway
manel.zarrouk@insight-centre.org

***Abstract*—Knowledge base creation and population are an essential formal backbone for a variety of intelligent applications, decision support and expert systems and intelligent search. While the abundance of unstructured text helps in easing the knowledge acquisition gap, the ambiguous nature of language tends to impact accuracy when engaging in more complex semantic analysis. Controlled Natural Languages (CNLs) are subsets of natural language that are restricted grammatically in order to reduce or eliminate ambiguity for the purposes of machine processability, or unambiguous human communication within a domain or industry context, such as Simplified English. This type of human-oriented CNL is under-researched despite having found favor within industry over many years. We describe a novel dataset which aligns a representative sample of Simplified English Wikipedia sentences with a well known machine-oriented CNL. This linguistic resource is both human-readable and semantically machine interpretable and can benefit a variety of NLP and knowledge based applications.**

***Index Terms*—Natural Language Processing, Controlled Natural Language, Knowledge Extraction, Semantic Web**

I. Introduction

Knowledge base creation and population is of paramount importance for a variety of applications including decision support and expert systems. Formal knowledge may also act as semantic backbone for language processing and information retrieval applications. A knowledge base for a particular domain may be either non existent or incomplete. While the abundance of unstructured text offers a means of easing the knowledge acquisition gap, the highly ambiguous nature of language impacts on accuracy when moving beyond the entity level to more complex semantic interpretation of text for knowledge creation. Controlled Natural Languages (CNLs) are unambiguous natural languages based on a restricted grammar that map into formal knowledge structures [1]. CNLs are an attempt to circumvent linguistic ambiguity and have found applications with respect to formal knowledge creation, ontology development and domain specific knowledge based machine translation [2]. Aside from specific knowledge gathering activities, which by the vary nature of the task, demand a restricted natural language interface for knowledge creation by domain experts, CNLs offer little incentive to the average user to create formal knowledge, even as implicit side effect of an authoring effort involving semantic annotation or enrichment of text content. A subcategory of CNL which offers a middle ground of reduced ambiguity for semantic interpretation but less restriction than a machine- oriented CNL. Human-oriented CNLs [3] have been in wide spread use for many years. Their development was motivated by the purposes of language learning, and unambiguous communication between humans in a domain context. An example of human-oriented CNL is the Simplified text such as Simple Wikipedia[1]. It is a form of text written using style guides[2] to reduce complexity and ambiguity of the language, especially for non-native speakers and juniors. Some of the written style guides recommended for authors in Simple Wikipedia include: *use active voices, avoid compound sentences(e.g conjunctions), avoid idioms (multi-words), keep sentences short and informative*. Other human-oriented CNLs include ASD Simplified Technical English[3], developed to improve the readability and comprehensibility in technical documents. In addition to, Boeing Technical English to improve the communication between people for air traffic control [4]. The development and planning of these CNLs appears often community driven, many of which at first glance, like Simplified Wikipedia, may been inspired by Basic English [5]. Human-oriented CNLs unlike machine-oriented CNLs [3] **do not** attempt to unambiguously map into formal knowledge structures, as the communication goal is human to human and not human to machine. An interesting premise is how close linguistically such languages are to machine-oriented CNLs and if so to what degree? Moreover, rewriting all or most of a human-oriented CNLs into a machine-oriented CNL which can be unambiguously parsed into a formal knowledge structure [17] could unlock significant silos of implicit general purpose domain knowledge, contained within existing human-oriented CNL content.

This paper focuses on our initial experiments with respect to the computational linguistic analysis of a corpus of human-oriented CNL represented in Simplified English, as well as the

[1] https://simple.wikipedia.org/wiki/

[2] https://simple.wikipedia.org/wiki/Wikipedia:How_to_write_Simple_English_pages

[3] http://www.asd-ste100.org/

978-1-5386-7325-6/18/$31.00 ©2018 IEEE
DOI 10.1109/WI.2018.00-58

investigation of the feasibility of rewriting Simplified English into a well know machine-oriented CNL called ACE [7]. Since no ground truth exists, it has been necessary to engineer one for our experiments.

We present a linguistic resource that is both human-readable and semantically machine interpretable. This resource is a snapshot taken from the abstracts of the Simple English Wikipedia dump[4]. The selected abstracts are rewritten into a machine-oriented CNL, by applying some rules on the syntactic structures of the sentences to be accepted and parsed by the CNL semantic parser. To our knowledge this resource is the first human to machine CNL aligned dataset. The paper is structured as follows: Section II describes related work, Section III presents the corpus collection, processing and analysis. In Section IV, we discuss the results and the evaluation, and finally, Section V offers a conclusion.

II. Related Work

Attempto Controlled English (ACE) is a CNL developed for knowledge representation. We choose ACE for our experiments for the following reasons: 1) It is a widely adopted CNL with an expressive grammar. 2) The language can be automatically mapped into different formal languages such as Discourse Representation Structures (variant of first order logic) [7] and subsequently a subset of ACE can be converted to the Web Ontology Language (OWL). 3) It also provides access to different tools and resources to use the language such as the ACE parsing engine - APE[5], which we use for our validation.

The work from [8] analysed the text in Simple Wikipedia and Wikipedia linguistic features to produce a new resource corpus that can help sentence simplification research. It provides a new simplification dataset that is an improved version over Simple Wikipedia. Their analysis over the text in both Simple Wikipedia and Wikipedia was done for the purpose of showing the amount of simplification between two parallel sentences from both Wikipedias, and introduce a new comparative approach to simplification corpus analysis. The main difference between our work is that we analyse both Wikipedias for the purpose of rewriting simplified English into a formal knowledge using CNLs as they tend to be less ambiguous.

In [9] the authors present a text rewriting approach to increase the amount of labeled data available for model training. They analysed Simple Wikipedia and Wikipedia parallel corpora to automatically extract rewrite rules, then generate multiple versions of sentences annotated with gold standard labels for the purposes of semantic role labeling. Our work is similar to them in that we also use rewrite rules to generate gold standard resource. However, the main difference is that we use rewrite rules to generate CNL sentences from Simple Wikipedia sentences, which means our efforts are directed towards a different purpose.

In [10] the author, presents a semantically annotated corpus developed using a bootstrapping approach using NLP techniques for parsing and semantic interpretation, together with an interface for collaborative annotation of experts and crowdsourcing community. Although they generated a semantic resource with deep semantics, the resource is small as it includes less than 5k sentences which is much smaller than ours.

The other resources of ProPBank [11] and Framenet [12] are semantically annotated corpora. Although, both of the resources are valuable to the community, they lack the deep semantic representation that combines different layers of semantic annotation into one formalism [10].

III. Corpus Analysis

We first collected a dataset of the abstracts from the Simple Wikipedia dump and its parallel Wikipedia dump. All experiments are performed on Simple Wikipedia abstracts only and their corresponding Wikipedia abstracts. Table I, describes the style guides from Simple Wikipedia on how to write simplified text versus several common stylistic properties observed across several CNLs (both human and machine oriented) [13]. From the table we can see that both texts significantly overlap in most of the properties.

A. Corpus Collection and Pre-processing

We collected the XML formatted dumps containing both Simple Wikipedia abstracts and their corresponding Wikipedia abstracts. Then, we converted the XML format into JSON[6] format. After that we cleaned the text using regular expressions which includes removing special characters, remove incomplete or blank sentences and abstracts, text between brackets, etc. In table II, we show all the steps performed for the preprocessing of the dumps.

The total number of abstracts before cleaning was around 74k in the Simple Wikipedia dump. Since we had not yet cleaned the dump, we did not count the parallel Wikipedia abstracts. After cleaning the data, we extracted around 48.8k abstracts from Simple Wikipedia and in parallel we found around 27.5k abstracts in Wikipedia as some of the abstracts are found blank, incomplete or missing. The total number of sentences extracted from the cleaned Simple Wikipedia was around 87k, including 968.2k tokens, with most of the abstracts including two sentences. In parallel, the total number of sentences extracted from the cleaned Wikipedia dump was around 39.2k, including around 586.7k tokens. In table III, we show a comparison between the two dumps before and after cleaning.

In order to test our intuition with respect to rewriting simplified text represented in the Simple Wikipedia abstracts to a machine-oriented CNL, we decided to analyze the presence of CNLs features in both within text abstracts of Simplified Wikipedia and the corresponding Wikipedia abstracts, the purpose being to measure how similar the Simplified text

[4] https://dumps.wikimedia.org/simplewiki/

[5] https://github.com/Attempto/APE

[6] Java Script Object Notation

TABLE I: Comparison showing the overlap between CNL rules and Simplified Text rules

Metric	Common CNL Rules	Simplified Text Rules
Short sentences	should not exceed 20 tokens	Keep sentences short and informative
Use active voices	recommended	recommended
Lexicon	Use approved words from the Dictionary (controlled lexicon)	Basic English Word-list
Idioms	Do not make noun clusters of more than three nouns	Avoid idioms (multi-words)

TABLE II: Corpus collection and Pre-processing steps

Metric	Simple Wikipedia	Wikipedia
Corpus	Abstracts of all articles (extracted from the dump)	Abstracts of parallel Simple Wikipedia articles
Format	XML dump converted to JSON format	
Pre-processing	Text cleansing using regular expressions (e.g. remove special characters, very short sentences, blank abstracts, text between brackets..etc)	
	Split each abstract into sentences (i.e sentence segmentation)	
	Split each sentence into tokens (i.e tokenization)	
	Run Part of Speech tagging (POS) using NLTK tagger using Penn Treebank Tag Set [16] over each sentence and and create a list of POS tags for each sentence in the corpus (POS structures list).	

TABLE III: Comparison of Simple Wikipedia & parallel Wikipedia abstracts

Metric	Simple Wikipedia	Parallel Wikipedia
No. of abstracts before cleaning	74,067	N/A
No. of abstracts after cleaning	48,880	27,539
Total No. of sentences in the corpus	87,088	39,252
Total No. of tokens in the corpus	968,231	586,732

to the common CNL features identified in [13]. In regard to the feasibility of rewriting Simplified Wikipedia sentences to a CNL, our first assumption is that the simplified text should be logically less ambiguous and less complex than standard Wikipedia unstructured text. Consequently, its linguistic properties will overlap significantly more with CNLs than unstructured text. So in order to measure this, we analyzed some of the measurable properties from [13] that should be present in the Simplified English text, given that it could be rewritten into CNL. As shown in table IV, the first metric is the length of sentences (number of tokens/sentence). In Simple Wikipedia we found that more than 90% of the sentences does not exceed the 20 tokens. On the other side, sentences from the parallel Wikipedia abstracts are usually exceeding this limit. Although the guidelines for writing Simple Wikipedia abstracts recommended the authors to avoid using passive voices, we found that 34% of the sentences did not follow this rule. On the other hand, 51% of the sentences in Wikipedia articles, are written using the passive voice. Gerunds are the words that are formed with verbs but act as nouns e.g go swimming, were found to be 6% in Simple Wikipedia sentences and 21% in Wikipedia sentences. CNLs usually use determiners before nouns, so our test found that the tags which preceded nouns in the Simple Wikipedia sentences are ranked as follows: 1) Determiners, 2) Noun Phrases, 3) Prepositions, 4) Adjectives, but in the Wikipedia sentences the list was different as follows 1) Nouns, 2) Noun Phrases, 3) Prepositions, 4) Determiners. Moreover, noun clusters are found in 4% of the Simple Wikipedia sentences and 8% in the Wikipedia sentences. Hence, based on the observations above, we can confirm our hypothesis that the linguistic properties of Simplified Wikipedia text overlap more with CNLs than unstructured text.

Based on the results above, we conducted a deeper analysis of the Simple Wikipedia POS structures which overlapped completely with the CNL rules. This meant extracting all abstracts which follow the common CNLs rules in table IV from the original dataset dump, excluding the remainder. As shown in table V, the total number of abstracts from the Simple Wikipedia dump that follow the CNL rules in Table IV are found to be around 20.4k. These abstracts include around 36.5k sentences, with 383.5k tokens. The result is our

TABLE IV: Results of analysing the CNL properties across Simple Wikipedia and Wikipedia sentences

Metric	Simple Wikipedia	Parallel Wikipedia
Maximum Tokens/sentence ≤20	Yes	No
Passive voices	34%	51%
Gerunds	6%	21%
Articles preceding nouns	DT,NNP,IN,JJ	NN,NP,IN,DT
Noun clusters	4%	8%

TABLE V: Results of extracting the Simple Wikipedia abstracts that follow the CNL rules.

Metric	Result	Percentage from the total corpus
No. of abstracts that are fully overlapping with the common CNL properties.	20,647	42.2%
Total No. of Sentences	36,560	42%
Total No. of Tokens	383,555	39.6%

reference corpus for our remaining experiments.

IV. Results

Since Simplified English rules and style guides request authors to use preferred sentence forms such as `Subject-Verb-DirectObect`, and `Subject-Verb-IndirectObject`[7]), our second hypothesis is that most of the sentences, which have the same POS tags structure/pattern (forms), can be rewritten into CNL using a few rules without changing their semantics. So, we created a dictionary that includes the POS tags structures of all the sentences in the corpus. The main aim of creating this dictionary is to discover to which extent the authors of the Simple Wikipedia abstracts followed the guidelines of writing preferred sentence structures for writing the simplified text. So, we grouped all similar POS tags structures together and count them, in order to estimate the percentage of unique POS tags structures and the percentage of repeated POS structures in the corpus. This would thus help us approximate the number of rewrite rules needed to map Simplified Wikipedia sentences into CNL. We found total number of 22,083 unique POS tag structures. Next, we estimated the number of sentences that belongs to a group of POS tags structures. As shown from table VI, we present 5 cases from the dictionary. For example, the dictionary shows that 102 POS tags structures include around 7.8k sentences, and 629 POS tags structures include around 12.6k sentences. This analysis indicated that there are a lot of repeated POS structures that could be found in the corpus.

We extracted the top 5 repeated POS tags structures in the corpus and we show our analysis on them in table VII. We found that the first structure is found 634 times in the corpus. We experimented with rewriting a few sentences which match this POS structure into ACE. So, in order to rewrite this POS tags structure, we need to chunk the noun clusters into one noun, to be accepted by the APE parser and validated as an ACE sentence for automatic translation into DRS formalisms. Although POS tags structures number 2, 3 and 5 are not as frequent as POS tags structure number 1, they did not require any rewriting to be accepted by the ACE parser or converted into DRS as they tend to be very basic sentences. The POS tags structure number 4, can rewritten using a noun phrase chunker in the beginning of the sentence to chunk the names as one noun, and another noun phrase chunker in the end of the sentence to chunk the combination consists of a noun after an adjective into one noun. Although table VI shows that there are some repeated POS tags structures in the corpus, still the number of sentences to POS tags structures ratio (12,630 sentence /630 POS tags structure) is not high, if we compared it to the total size of the corpus which is around 36.5k sentences. This means that in order to rewrite 12,630 sentences into CNL we need to implement rules that can cover the 630 different POS tags structures, which is a lot of rules that will lead to rewriting only 35% of the corpus. Although it was shown from table VII that some few basic sentences did not require a rule for rewriting these sentences, the table showed that the percentage of these is still very low.

A. Annotation and Validation

So, in order to overcome this problem we found from our analysis that a common rule from table VII contains noun clusters and/or adjective noun clusters - noun phrase chunks effectively. So, we applied chunking rules on all the sentences in the corpus and assigned them a new POS tag `COMP`. A new set of new POS tags structures are created and the results from this analysis is summarized in table VIII. In the table we notice that around 19.2k sentences are captured together under 300 POS tags structures only, out of total 13,867 POS tags structures. This means that if we can rewrite the 300 structures into ACE CNL, we envisage having around 19.2k sentences in ACE CNL format and consequently into the DRS formal representation and exported to OWL.

Open further analysis after extracting from the corpus the 19.2k sentences that belong to the 300 POS tags structures,

[7] https://simple.wikipedia.org/wiki/Wikipedia:How_to_write_Simple_English_pages

TABLE VI: Grouping sentences that belong to the same POS tags structure

No. of POS tags Structures	Percentage from the total No. of POS tags Structures	No. of sentences	Percentage from the total No. of sentences
102	0.4%	7,809	21.3%
115	0.5%	8,080	22.1%
182	0.8%	9,203	25.1%
289	1.3%	10,460	28.6%
629	2.8%	12,630	34.5%

TABLE VII: A table showing the sentences that belong to the top 5 dominating POS tags structures after rewriting into a CNL.

Rank	POS Pattern/Structure	Count	Examples	ACE CNL
1	NNP VBZ DT NN IN NNP IN DT NNP NNPS	634	• Macon is a city of Illinois in the United States. • Agency is a city of Iowa in the United States	• Macon is a city of Illinois in the n:United-States. • Agency is a city of Iowa in the n:United-States.
2	NNP VBZ DT NN IN NNP	295	• Aalborg is a city in Denmark • Helene is a moon of Saturn	• Aalborg is a city in Denmark. • Helene is a moon of Saturn.
3	NNP VBZ DT NN	199	• Thirteen is a number • Waitby has a castle	• Thirteen is a number • Waitby has a castle
4	NNP NNP NNP VBD DT JJ NN	196	• Guy Henry Ourisson was a French chemist. • Edna May Oliver was an American actress	• Guy-Henry-Ourisson is a n:French chemist. • Edna-May-Oliver is an n:American-actress.
5	NNP VBZ DT JJ NN	149	• Adelite is a rare mineral. • Zugzwang is a chess term.	• Adelite is a rare mineral. • Zugzwang is a chess term.

TABLE VIII: Grouping sentences that belong to the same POS tags structure after chunking nouns and adjectives.

No. of POS tags Structures	Percentage from the total No. of POS tags Structures	No. of sentences	Percentage from the total No. of sentences
300	2.1%	19,236	52.6%
605	4.3%	21,203	58%

we applied a total of 10 rules[8]) to the sentences, where some of the results of these rules can be shown in table IX. The table shows the top 5 POS tags structures after the chunking took place. The highest occurring POS tags structure was found 2664 times. The rewriting rule for this structure is to chunk the nouns at the end of the sentence. The second most dominating structure was found 1399 times, and the rewriting rule chunked the nouns at the beginning and of the sentence. Structure number 3 occurred 1088 times, and the rewriting rule chunked the nouns at the end of the sentence. Structure number 4 occurred 926 times and the rewriting rule chunked a noun/group of nouns, followed by POS tag IN, followed by a noun/group of nouns. The last structure occurred 885 times, where the past verb of the sentence is rewritten into the present tense and the nouns at the beginning and end of the sentence are chunked.

B. Evaluation

The evaluation of the generated resource is performed on two sides, 1) The system coverage, and 2) The Semantic similarity of the mapped sentences. Since we developed syntactic rules to cover the top 300 POS tags structures, the system coverage is to estimate how many sentences the system was

[8] https://drive.google.com/open?id=1eBUXZ8tESIML3jEptqG4aci4-_BmODkP

TABLE IX: A table showing the sentences that belong to the top 5 dominating POS tags structures (using chunking) after rewriting into a CNL.

Rank	POS Pattern/Structure	Count	Examples	ACE CNL
1	COMP VBZ DT COMP IN COMP	2664	• Alkmene is a person in Greek mythology. • Anyang is a city in South Korea.	• Alkmene is a person in the n:Greek mythology. • Anyang is a city in the n:South-Korea.
2	COMP VBZ DT COMP	1399	• Alicia Bridges is an American singer. • Blood transfusion is a medical term.	• Alicia-Bridges is an n:American-singer. • Blood-transfusion is a n:medical-term.
3	COMP VBZ DT COMP IN COMP IN DT COMP	1088	• Gilbert is a city of Iowa in the United States. • Iphiclides is a genus of Butterflies in the family Papilionidae.	• Gilbert is a city of Iowa in the n:United-States. • Iphiclides is a genus of Butterflies in the n:family-Papilionidae.
4	COMP VBZ DT COMP IN DT COMP IN COMP	926	• Anhui is a province in the east of China. • Agriculture is an important part of the economy of Azerbaijan.	• Anhui is a province in the n:east-of-China. • Agriculture is an important part of the n:economy-of-Azerbaijan.
5	COMP VBD DT COMP	885	• Abel Ricardo Laudonio was an Argentine boxer. • Braniff International Airways was an American airline.	• Abel-Ricardo-Laudonio is an n:Argentine-boxer. • Braniff-International-Airways is an n:American-airline.

TABLE X: Evaluation of the coverage of the rewriting system.

Metric	Total No. of sentences	Coverage
Rewritten into ACE CNL	17,199	89.4%
Not rewritten into ACE CNL	2,037	10.6%

able to rewrite out of the whole corpus. After we applied the rules on the corpus, we can see from Table X that we were able to rewrite 17,199 sentences out of 19,236 sentences to the ACE CNL. Thus, the system coverage is equal to 89.4%, and sentences that failed to parse are 2037 sentences that represent 10.6% of the corpus. These sentences failed to parse as the developed rules failed to rewrite them into an interpretable ACE CNL structure, thus they were refused by the APE parser. The rewrite rules are developed for the most repeated POS structures to rewrite as much sentences as possible, neglecting individual POS structures. All the rewritten sentences are passed to the APE parser web service and validated for being successfully parsed.

The second evaluation is done to ensure that the semantics has not changed after the rewriting happened, we need to test whether the generated ACE CNL sentence preserved the same semantics of its mapped SE sentence. So, we extracted a representative sample from the corpus after estimating the sample size from [18] and it was found to be 527 sentences. Then, we computed the semantic textual similarity between the SE sentence and its mapped ACE CNL sentence based on the research in [19]. The result showed that our system preserves the semantic similarity between all the SE sentences, and their mapped ACE CNL sentences in the sample set. The resource is available for download[9]) and includes all the 17,199 sentences from original Simple Wikipedia and their parallel ACE rewritten sentences. These ACE sentences should be a very valuable resource to the community for exploitation in different applications related to NLP for knowledge base population.

V. Conclusion

We have presented a linguistic resource that is both human-readable and semantically machine interpretable. To our knowledge this resource is the first aligned dataset across a human-oriented and machine-oriented CNL as well as unstructured text. The dataset is well represented in that it takes almost the entire Simplified Wikipedia abstract population post-cleansing. We have provided corpus statistics and linguistics analysis, which have confirmed our hypotheses with respect to rewriting Simplified English to ACE and its subsequent transformation into logical and knowledge representations such as DRS and OWL respectively.

This resource could be exploited in other fields beyond CNLs for knowledge based population. Potential applications include generating general knowledge for expert and knowledge based systems and ontology aware NLP applications

[9] https://drive.google.com/open?id=1eBUXZ8tESIML3jEptqG4aci4-_BmODkP

as well as knowledge based MT, automated reasoning, language learning as well as teaching and learning resources for knowledge engineering and logic programming. In addition to, generalization to other languages can be done using present frameworks such as GF to translate ACE to other languages [20].

Future work with respect to this resources will involve augmenting the aligned human to machine-oriented CNL content with semantic metadata such as RDF[10] generated from ACE2OWL[11] [17]. Other work will involve additional corpus analysis and rule generation for rewriting less common Simplified English POS patterns.

Acknowledgment

This work has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 and in part by the SSIX Horizon 2020 project (grant agreement No 645425).

References

[1] Hazem Safwat and Brian Davis. 2014. *A brief state of the art of CNLs for ontology authoring*. International Workshop on Controlled Natural Language. pp. 190-200. Springer.

[2] Tobias Kuhn. 2014. *A survey and classification of controlled natural languages*. Computational Linguistics 40, pp. 121-170.

[3] Hujisen WO. 1998. *Controlled language: an introduction*. Second International Workshop on Controlled Language Applications (CLAW), Pittsburgh.

[4] Wojcik, Richard H., Heather Holmback, and James Hoard. 1998. *Boeing Technical English: An extension of AECMA SE beyond the aircraft maintenance domain*. Second International Workshop on Controlled Language Applications (CLAW), Pittsburgh.

[5] Ogden, Charles Kay. 1944. *Basic English: A general introduction with rules and grammar*. Vol. 29. K. Paul, Trench, Trubner.

[6] Hazem Safwat and Brian Davis. 2017. *CNLs for the semantic web: a state of the art*. Language Resources and Evaluation, 51(1), pp.191-220.

[7] Fuchs, Norbert E., Kaarel Kaljurand, and Tobias Kuhn. 2008. *Attempto controlled english for knowledge representation*. Reasoning Web (pp. 104-124). Springer, Berlin, Heidelberg.

[8] Xu, Wei, Chris Callison-Burch, and Courtney Napoles. 2015. *Problems in current text simplification research: New data can help*. Transactions of the Association of Computational Linguistics, 3(1), pp.283-297.

[9] Woodsend, Kristian, and Mirella Lapata. 2017. *Text rewriting improves semantic role labeling*. Journal of Artificial Intelligence Research, 51, pp.133-164.

[10] Basile, Valerio, Johan Bos, Kilian Evang, and Noortje Venhuizen. 2012. *Developing a large semantically annotated corpus*. Eighth International Conference on Language Resources and Evaluation.

[11] Martha Palmer, Paul Kingsbury, and Daniel Gildea. 2005. *The proposition bank: An annotated corpus of semantic roles*. Computational Linguistics, 31(1):71106.

[12] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. *The Berkeley FrameNet project*. 17th International Conference on Computational Linguistics.

[13] OBrien, Sharon. 2003. *Controlling controlled english. an analysis of several controlled language rule sets*. Proceedings of EAMT-CLAW, 3, pp.105-114.

[14] Bird, Steven, and Edward Loper. 2004. *Text rewriting improves semantic role labeling*. Journal of Artificial Intelligence Research, 51, pp.133-164.

[15] Marcus, Mitchell P., Mary Ann Marcinkiewicz, and Beatrice Santorini. 1998. *Building a large annotated corpus of English: The Penn Treebank.*. Computational linguistics, 19(2), pp.313-330.

[16] Fuchs, Norbert E., Kaarel Kaljurand, and Tobias Kuhn. 2008. *Discourse representation structures for ACE 6.0.* Department of Informatics.

[17] Kaljurand, Kaarel, and Norbert E. Fuchs. 2006. *Mapping Attempto Controlled English to OWL DL.* 3rd European Semantic Web Conference. Demo and Poster Session, Budva, Montenegro.

[18] Smith, S. 2013. *Determining sample size: How to ensure you get the correct sample size.* E-Book (c) Qualtrics Online Sample.

[19] Han, L., Kashyap, A.L., Finin, T., Mayfield, J. and Weese, J. 2013. *UMBC_EBIQUITY-CORE: semantic textual similarity systems.* In Second Joint Conference on Lexical and Computational Semantics (SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity (Vol. 1, pp. 44-52).

[20] Kaljurand, K. and Kuhn, T. 2013. *A multilingual semantic wiki based on Attempto Controlled English and Grammatical Framework.* In Extended Semantic Web Conference (pp. 427-441). Springer, Berlin, Heidelberg.

[10] https://www.w3.org/RDF/

[11] https://www.w3.org/2001/sw/wiki/ACE