From Genesis to Creole language: Transfer Learning for Singlish Universal Dependencies Parsing and POS Tagging

HONGMIN WANG, University of California Santa Barbara, USA
JIE YANG, Singapore University of Technology and Design, Singapore
YUE ZHANG, West Lake University, Institute for Advanced Study, China

Singlish can be interesting to the computational linguistics community both linguistically as a major low-resource creole based on English, and computationally for information extraction and sentiment analysis of regional social media. In our conference paper, Wang et al. [2017], we investigated part-of-speech (POS) tagging and dependency parsing for Singlish by constructing a treebank under the Universal Dependencies scheme, and successfully used neural stacking models to integrate English syntactic knowledge for boosting Singlish POS tagging and dependency parsing, achieving the state-of-the-art accuracies of 89.50% and 84.47% for Singlish POS tagging and dependency respectively. In this work, we substantially extend Wang et al. [2017] by enlarging the Singlish treebank to more than triple the size and with much more diversity in topics, as well as further exploring neural multi-task models for integrating English syntactic knowledge. Results show that the enlarged treebank has achieved significant relative error reduction of 45.8% and 15.5% on the base model, 27% and 10% on the neural multi-task model, and 21% and 15% on the neural stacking model for POS tagging and dependency parsing respectively. Moreover, the state-of-the-art Singlish POS tagging and dependency parsing accuracies have been improved to 91.16% and 85.57% respectively. We make our treebanks and models available for further research.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; **Natural language processing**; **Language resources**;

Additional Key Words and Phrases: Dependency Parsing, Universal Dependencies, Part-of-speech Tagging, Transfer Learning, Neural Stacking, Multi-task Network, Creole Language, Singlish

ACM Reference Format:
Hongmin Wang, Jie Yang, and Yue Zhang. 2010. From Genesis to Creole language: Transfer Learning for Singlish Universal Dependencies Parsing and POS Tagging. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 9, 4 (March 2010), 23 pages. https://doi.org/0000001.0000001

1 INTRODUCTION

Languages evolve temporally and geographically, both in vocabulary as well as in syntactic structures. When major languages such as English or French are adopted in another culture as the primary language, they often mix with existing languages or dialects in that culture and evolve into a stable language called a creole. Examples of creoles include the French-based Haitian Creole, and Colloquial Singaporean English (Singlish) [Mian-Lian and Platt 1993], an English-based creole.

Authors' addresses: Hongmin Wang, University of California Santa Barbara, 2104 Harold Frank Hall, Santa Barbara, CA, 93106-5110, USA, hongmin_wang@cs.ucsb.edu; Jie Yang, Singapore University of Technology and Design, 8 Somapah Rd, 487372, Singapore, jieynlp@gmail.com; Yue Zhang, West Lake University, Institute for Advanced Study, 310024, 18 Shilongshan Road, Hangzhou, Zhejiang, China, yue_zhang@sutd.edu.sg.

2375-4699/2010/3-ART $15.00
https://doi.org/0000001.0000001

While the majority of the natural language processing (NLP) research attention has been focused on the major languages, little work has been done on adapting the components to creoles. One notable body of work originated from the featured translation task of the EMNLP 2011 Workshop on Statistical Machine Translation (WMT11) to translate Haitian Creole SMS messages sent during the 2010 Haitian earthquake. This work highlights the importance of NLP tools on creoles in crisis situations for emergency relief [Hewavitharana et al. 2011; Hu et al. 2011].

Singlish is one of the major languages in Singapore, with borrowed vocabulary and grammars[1] from a number of languages including Malay, Tamil, and Chinese dialects such as Hokkien, Cantonese and Teochew [Leimgruber 2009, 2011], and it has been increasingly used in written forms on web media. However, fluent English speakers unfamiliar with Singlish would find the creole hard to comprehend [Harada 2009]. Correspondingly, fundamental English NLP components such as POS taggers and dependency parsers perform poorly on such Singlish texts as shown in Table 2 and 4. One other example is that Seah et al. [2015] adapted the Socher et al. [2013] sentiment analysis engine to the Singlish vocabulary, but failed to adapt the parser. Since dependency parsers are important for tasks such as information extraction [Miwa and Bansal 2016] and discourse parsing [Li et al. 2015], this hinders the development of such downstream applications for Singlish in written forms and thus makes it crucial to build a dependency parser that can perform well natively on Singlish.

To address this issue, we took the first attempt starting with investigating the linguistic characteristics of Singlish and specifically the causes of difficulties for understanding Singlish with English syntax. We found that, despite the obvious attribute of inheriting a large portion of basic vocabularies and grammars from English, Singlish not only imports terms from regional languages and dialects, its lexical semantics and syntax also deviate significantly from English [Leimgruber 2009, 2011]. We further categorized the challenges and formalized their interpretation using Universal Dependencies [Nivre et al. 2016] and created a Singlish dependency treebank with 1,200 sentences.

Based on the intricate relationship between Singlish and English, we built a Singlish parser by leveraging knowledge of English syntax as a basis. This overall approach is illustrated in Figure 1 (a). In particular, we trained a basic Singlish parser with one of the best off-the-shelf neural dependency parsing model using biaffine attention [Dozat and Manning 2017], and improved it with knowledge transfer by adopting neural stacking [Chen et al. 2016; Zhang and Weiss 2016] to integrate the English syntax. Since POS tags are important features for dependency parsing [Chen and Manning 2014; Dyer et al. 2015], we also trained a POS tagger for Singlish following the same idea by integrating English POS knowledge using neural stacking.

Results by our conference paper show that English syntax knowledge brings 51.50% and 25.01% relative error reduction on POS tagging and dependency parsing respectively, resulting in a Singlish dependency parser with 84.47% unlabeled attachment score (UAS) and 77.76% labeled attachment score (LAS).

In this paper, we further extend the work by substantially enlarging the Singlish dependency treebank to more than triple the size, using data with much more diversity in terms of topics and time span from multiple local Internet forums. We show that this leads to more substantiations in Singlish syntactic constructions as illustrations by case analyzes in section 3.4 and a summarization of all Singlish terms and their meanings in Tables **??** to **??** in Appendix **??**. Furthermore, we explore neural multi-task models for transferring English knowledge to Singlish dependency parsing, as shown in Figure 1 (b), which is done by joint training of a Singlish and an English parser with a

[1]We follow Leimgruber [2011] in using "grammar" to describe "syntactic constructions" and we do not differentiate between the two expressions in this paper.

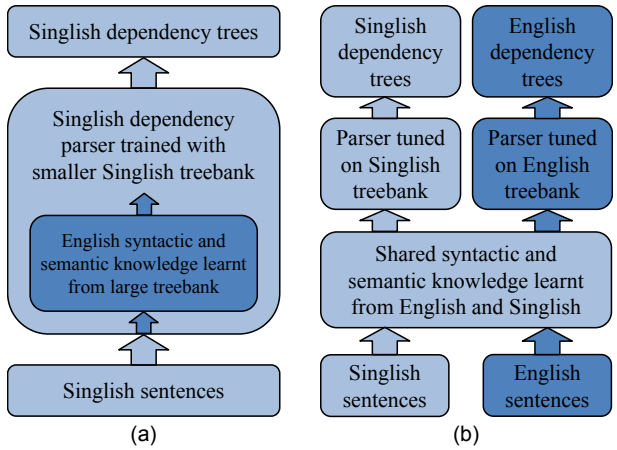

Fig. 1. Overall model diagrams: (a) Neural Stacking Model (b) Multi-task Model

shared module to enable knowledge sharing. The same has been applied to POS tagging and it has achieved significant improvement over the base models for both tasks.

As a significantly extended version of our conference work, we show in this paper that the current state-of-the-art Singlish dependency parsing has been improved to 85.57% UAS and 79.12% LAS and POS tagging accuracy has been improved from 89.50% to 91.16%. We make our Singlish dependency treebank, the source code for training a dependency parser and the trained model for the parser with the best performance freely available[2] to facilitate future research.

The paper is organized as follows: section 2 discusses the related works in the areas of transfer learning for cross-lingual parsing and neural stacking and multi-task models, section 3 demonstrates the adaption of Universal Dependencies to Singlish during the dataset construction and extension processes, section 4 describes the base, neural stacking, and multi-task models for POS tagging and experiment results using the original and extended treebanks, section 5 describes the different models for Singlish parsing followed by discussions on various experiments in section 6, and finally section 7 concludes the paper.

2 RELATED WORK

Neural networks have led to significant performance improvement for dependency parsing, including transition-based parsing [Andor et al. 2016; Ballesteros et al. 2015; Chen and Manning 2014; Dyer et al. 2015; Weiss et al. 2015; Zhou et al. 2015], and graph-based parsing [Dozat and Manning 2017; Kiperwasser and Goldberg 2016]. In particular, the biaffine attention method of Dozat and Manning [2017] uses deep bi-directional long short-term memory (bi-LSTM) networks for high-order non-linear feature extraction, producing one of the highest-performing graph-based English dependency parsers. We adopt this model as the basis for our Singlish parser.

Our work belongs to a line of work on transfer learning for parsing, which leverages English resources in Universal Dependencies to improve the parsing accuracies of low-resource languages [Cohen and Smith 2009; Ganchev et al. 2009; Hwa et al. 2005]. Seminal work employed statistical models. McDonald et al. [2011] investigated delexicalized transfer, where word-based features are removed from a statistical model for English so that POS and dependency label knowledge can be utilized for training a model for low-resource language. Subsequent work considered syntactic

[2] https://github.com/wanghm92/Sing_Par

similarities between languages for better feature transfer [Naseem et al. 2012; Täckström et al. 2012; Zhang and Barzilay 2015].

Recently, a line of work leverages neural network models for multi-lingual parsing [Ammar et al. 2016; Duong et al. 2015; Guo et al. 2015]. The basic idea is to map the word embedding spaces between different languages into the same vector space, by using sentence-aligned bilingual data. This gives consistency in tokens, POS and dependency labels thanks to the availability of Universal Dependencies [Nivre et al. 2016]. Our work is similar to these methods in using a neural network model for knowledge sharing between different languages. However, ours is different in the use of a neural stacking model and a multi-task learning model, which respect the distributional differences between Singlish and English words. This empirically gives higher accuracies for Singlish.

Neural stacking and neural multi-task models were previously used for cross-annotation [Chen et al. 2016] and cross-task [Zhang and Weiss 2016] joint-modeling on monolingual treebanks. We were the first to employ neural stacking on cross-lingual feature transfer from resource-rich languages to improve dependency parsing for low-resource languages and we further explore neural multi-task models on this task in this work. Besides these three dimensions in dealing with heterogeneous text data, another popular area of research is on the topic of domain adaption, which is commonly associated with cross-lingual problems [Nivre et al. 2007]. While this large strand of work is remotely related to ours, we do not describe them in details.

Unsupervised rule-based approaches also offer an competitive alternative for cross-lingual dependency parsing [Gelling et al. 2012; Gillenwater et al. 2010; Martínez Alonso et al. 2017; Naseem et al. 2010; Søgaard 2012a,b], and recently been benchmarked for the Universal Dependencies formalism by exploiting the linguistic constraints in the Universal Dependencies to improve the robustness against error propagation and domain adaption [Martínez Alonso et al. 2017]. However, we choose a data-driven supervised approach given the relatively higher parsing accuracy owing to the availability of resourceful treebanks from the Universal Dependencies project.

3 SINGLISH DEPENDENCY TREEBANK

3.1 Universal Dependencies for Singlish

Since English is the major genesis of Singlish, we choose English as the source of lexical feature transfer to assist Singlish dependency parsing. Universal Dependencies provides a set of multilingual treebanks with cross-lingually consistent dependency-based lexicalist annotations, designed to aid development and evaluation for cross-lingual systems, such as multilingual parsers [Nivre et al. 2016]. The current version of Universal Dependencies comprises not only major treebanks for 76 languages but also their siblings for domain-specific corpora and dialects. With the aligned initiatives for creating transfer-learning-friendly treebanks, we adopt the Universal Dependencies protocol for constructing the Singlish dependency treebank, both as a new resource for the low-resource languages and to facilitate knowledge transfer from English.

On top of the general Universal Dependencies guidelines, English-specific dependency relation definitions including additional subtypes are employed as the default standards for annotating the Singlish dependency treebank, unless augmented or redefined when necessary. The latest English corpus in Universal Dependencies v2.3[3] has 6 treebanks from different domains and the main collection is constructed from the English Web Treebank [Bies et al. 2012], comprising of web media texts such as blogs and online reviews, which potentially smooths the knowledge transfer to our target Singlish texts in similar domains. However, to be consistent with our previous work in terms of treebank extension and fair comparison for the experiments, we still adopt the annotation

[3] As of 16 April 2019, http://universaldependencies.org/

	UD English		Singlish		Extension	
	Sentences	Words	Sentences	Words	Sentences	Words
Train	12,543	204,586	900	8,221	3,050	27,368
Dev	2,002	25,148	150	1,384	150	1,384
Test	2,077	25,096	150	1,381	150	1,381

Table 1. Division of training, development, and test sets for the Singlish Treebank STB-ACL

standards compatible with the v1.4 Universal Dependencies English treebanks[4] and leave the version conversion to future works. The statistics of this dataset, from which we obtain English syntactic knowledge, is shown in Table 1 and we refer to this corpus as UD-Eng. This corpus uses 47 dependency relations and we show below how to conform to the same standard while adapting to unique Singlish grammars.

3.2 Challenges and Solutions for Annotating Singlish

The deviations of Singlish from English come from both the lexical and the grammatical levels [Leimgruber 2009, 2011], which bring challenges for analysis on Singlish using English NLP tools. The former involves imported vocabularies from the first languages of the local people and the latter can be represented by a set of relatively localized features which collectively form 5 unique grammars of Singlish according to Leimgruber [2011]. We find empirically that all these deviations can be accommodated by applying the existing English dependency relation definitions while ensuring consistency with the annotations in other non-English UD treebanks, which are explained with examples as follows.

Imported vocabulary: Singlish borrows a number of words and expressions from its non-English origins [Leimgruber 2009, 2011], such as "*Kiasu*" which means "*very anxious not to miss an opportunity*" in Hokkien[5], and "*makan*" of (1) in Figure 2 which means "*eat*" in Malay. These imported terms often constitute out-of-vocabulary (OOV) words with respect to a standard English treebank and result in difficulties for using English-trained tools on Singlish. All borrowed words are annotated based on their usages in Singlish, which mainly inherit the POS from their genesis languages. Table **??** and **??** in Appendix **??** summarizes all borrowed terms in the first version of the Singlish treebank. Table **??** to **??** in Appendix **??** summarizes all borrowed terms as well as their meanings looked up from the Singlish resources listed in section 3.3 and the *Urban Dictionary*[6].

Topic-prominence: This type of sentences start with establishing its topic, which often serves as the default one that the rest of the sentence refers to, and they typically employ an object-subject-verb sentence structure [Leimgruber 2009, 2011]. In particular, three subtypes of topic-prominence are observed in the Singlish dependency treebank and their annotations are addressed as follows:

First, topics framed as clausal arguments at the beginning of the sentence are labeled as "*csubj*" (clausal subject), as shown by "*Drive this car*" of (2) and "*Earn some kopi money*" of (3) in Figure 2, which is consistent with the dependency relations in its Chinese translation.

Second, noun phrases used to modify the predicate with the absence of a preposition is regarded as a "*nsubj*" (nominal subject). Similarly, this is a common order of words used in Chinese and one example is the "*SG*" of (4) in Figure 2.

Third, prepositional phrases moved in front are still treated as "*nmod*" (nominal modifier) of their intended heads, following the exact definition but as a Singlish-specific form of exemplification, as shown by the "*Inside tent*" of (5) in Figure 2.

[4]available at https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1827
[5]Definition by the *Oxford living Dictionaries for English.*
[6]https://www.urbandictionary.com/

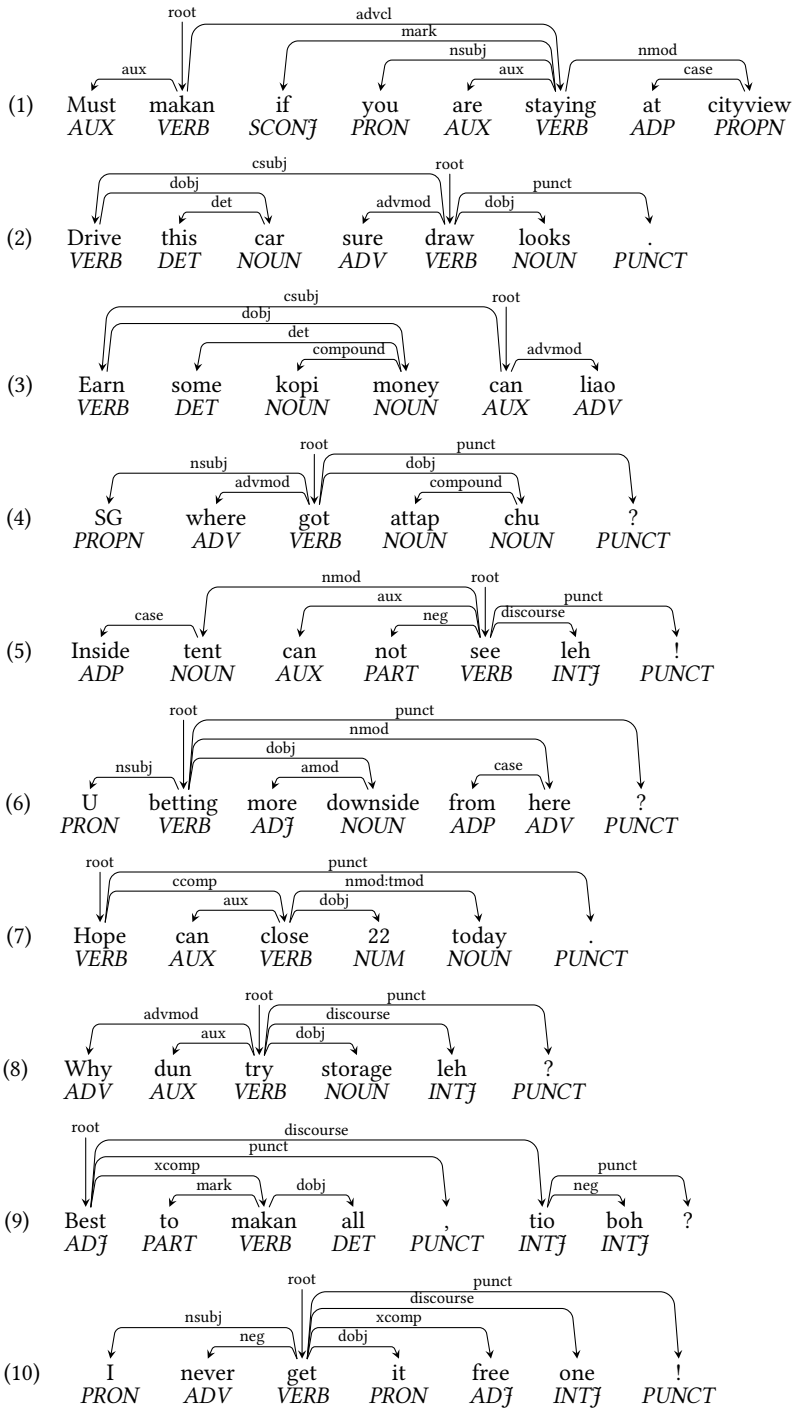

Fig. 2. Unique Singlish grammars. Arcs represent dependencies, pointing from the head to the dependent and the label on top. POS tags are below the words. (English translations available at Appendix **??**)

Although the "*dislocated*" (dislocated elements) relation in UD is also used for preposed elements, but it captures the ones "*that do not fulfill the usual core grammatical relations of a sentence*" and "*not for a topic-marked noun that is also the subject of the sentence*" [Nivre et al. 2016]. In these three scenarios, the topic words or phrases are in relatively closer grammatical relations to the predicate, as subjects or modifiers.

Copula deletion: Imported from the corresponding Chinese sentence structure, this copula verb is often optional and even deleted in Singlish, which is one of its diagnostic characteristics [Leimgruber 2009, 2011]. In UD-Eng standards, predicative "*be*" is the only verb used as a copula and it often depends on its complement to avoid copular head. This is explicitly designed in UD to promote parallelism for zero-copula phenomenon in languages such as Russian, Japanese, and Arabic. The deleted copula and its "*cop*" (copula) arcs are simply ignored, as shown by (6) in Figure 2. Copula deletion is also broadly shared in social media texts such as Tweets [Liu et al. 2018; O'Connor et al. 2018; Sanguinetti et al. 2018, 2017], which arises along with the prosperity of social media in recent two decades. The purpose of dropping components in social text media is mainly for conciseness of expression. However, the copula deletion phenomenon in Singlish mainly results from environmental influence [Leimgruber 2009], such as mixing Chinese dialects with English, which has been prevailing throughout its history.

NP deletion: Noun-phrase (NP) deletion often results in null subjects or objects. It may be regarded as a branch of "*Topic-prominence*" but is a distinctive feature of Singlish with a relatively high frequency of usage [Leimgruber 2011]. NP deletion is also common in pronoun-dropping languages such as Spanish and Italian, where the anaphora can be morphologically inferred. In one example, "*Vorrei ora entrare brevemente nel merito*"[7], from the Italian treebank in UD, "*Vorrei*" means "*I would like to*" and depends on the sentence root, "*entrare*", with the "*aux*"(auxiliary) relation, where the subject "*I*" is absent but implicitly understood. Similarly, we do not recover such relations since the deleted NP imposes negligible alteration to the dependency tree, as exemplified by the possibly missing "*it*" between "*Hope*" and "*can*" in (7) and the "*you*" between "*Why*" and "*dun*"(don't) in (8) in Figure 2.

Inversion: Inversion in Singlish involves either keeping the subject and verb in interrogative sentences in the same order as in statements, or tag questions in polar interrogatives [Leimgruber 2011]. The former also exists in non-English languages, such as Spanish and Italian, where the subject can prepose the verb in questions [Lahousse and Lamiroy 2012]. This simply involves a change of word orders and thus requires no special treatments. On the other hand, tag questions should be carefully analyzed in two scenarios. One type is in the form of "*isn't it?*" or "*haven't you?*", which are dependents of the sentence root with the "*parataxis*" relation.[8] The other type is exemplified as "*right?*", and its Singlish equivalent "*tio boh?*" (a transliteration from Hokkien) is labeled with the "*discourse*" (discourse element) relation with respect to the sentence root. See example (9) in Figure 2.

Discourse particles: Usage of clausal-final discourse particles, which originates from Hokkien and Cantonese, is one of the most typical features of Singlish [Leimgruber 2009, 2011; Lim 2007]. All discourse particles that appear in our treebank are summarized in Table **??** in Appendix **??** with the imported vocabulary. These words express the tone of the sentence and thus have the "*INTJ*" (interjection) POS tag and depend on the root of the sentence or clause labeled with "*discourse*", as is shown by the "*leh*" of (5) and (8) in Figure 2. The word "*one*" is a special instance of this type with the sole purpose being a tone marker in Singlish but not English, as shown by (10) in Figure 2.

[7]In English: (I) would now like to enter briefly on the merit (of the discussion).
[8]In UD: Relation between the main verb of a clause and other sentential elements, such as sentential parenthetical clause, or adjacent sentences without any explicit coordination or subordination.

3.3 Data Selection and Annotation

On top of the released first version of the Singlish treebank, we extend it to more than triple the size with much more diversity in terms of topics and time span.

Data Source: Singlish is used in written form mainly in social media and local Internet forums. After comparison,we first chose the *SG Talk Forum*[9] as the data source due to its relative abundance in Singlish contents. Specifically, they crawled[10] 84,459 posts using the *Scrapy* framework[11] from pages dated up to 25th December 2016, retaining sentences of length between 5 and 50, which total 58,310. Sentences were reversely sorted according to the log-likelihood of the sentence given by an English language model trained using the *KenLM* toolkit[12] [Heafield et al. 2013] normalized by the sentence length, so that those most different from standard English can be chosen. Among the top 10,000 sentences, 1,977 sentences contain unique Singlish vocabularies defined by *The Coxford Singlish Dictionary*[13], *A Dictionary of Singlish and Singapore English*[14], and the *Singlish Vocabulary* Wikipedia page[15]. The average normalized log likelihood of these 10,000 sentences is -5.81, and the same measure for all sentences in UD-Eng is -4.81. This means these sentences with Singlish contents are 10 times less probable expressed as standard English than the UD-Eng contents in the web domain. This contrast indicates the degree of lexical deviation of Singlish from English. Finally, 1,200 sentences were chosen from the first 10,000. More than 70% of the selected sentences are observed to consist of the Singlish grammars and imported vocabularies described in section 3.2. Thus the evaluations on this treebank can reflect the performance of various POS taggers and parsers on Singlish in general.

Annotation: The chosen texts are divided by random selection into training, development, and testing sets according to the proportion of sentences in the training, development, and test division for UD-Eng, as summarized in Table 1. The sentences are tokenized using the NLTK Tokenizer[16], and then annotated using the Dependency Viewer[17]. In total, all 17 UD-Eng POS tags and 41 out of the 47 UD-Eng dependency labels are present in the Singlish dependency treebank. Besides, 100 sentences are randomly selected and double annotated by one of the coauthors, and the inter-annotator agreement has a 97.76% accuracy on POS tagging and a 93.44% UAS and a 89.63% LAS for dependency parsing. A full summary of the numbers of occurrences of each POS tag and dependency label are included in Appendix **??**. We name the dataset **STB-ACL**.

Extension: With the success on improving the Singlish dependency parsing accuracy using the first released version of the treebank, we take a step further to extend the treebank comparable to the size of a UD treebank for the lower source languages. Specifically, we further apply the same procedures described above using all three major and popular forums, *SG Talk Forum*, *Hardwarezone*[18] and *SgForums*[19], by crawling all textual posts on all three forums dated until 19 June 2017. This brings two major differences in the source of data for the extended dataset:

- **Size**: The data source totals 420 million words which is more than 10 times larger.

[9] http://sgTalk.com
[10] Using *Python* package *beatifulsoup4* available at: https://pypi.python.org/pypi/beautifulsoup4
[11] https://scrapy.org/
[12] Trained using the *afp_eng* and *xin_eng* sources of English Gigaword Fifth Edition (*Gigaword*).
[13] http://72.5.72.93/html/lexec.php
[14] http://www.singlishdictionary.com
[15] https://en.wikipedia.org/wiki/Singlish_vocabulary
[16] http://www.nltk.org/api/nltk.tokenize.html
[17] http://nlp.nju.edu.cn/tanggc/tools/DependencyViewer.exe
[18] http://forums.hardwarezone.com.sg/
[19] http://sgforums.com/forums

- **Diversity**: The three forums cover 10 times more topics. In particular, *Hardwarezone* has more than 46 sub-forums cover topics including technology events, information technology diagnosis, digital entertainment lifestyle and etc, *SgForums* is a general open-domain forum focusing more on daily news and casual topics.

We remove about 1% contents written in non-English characters and then tokenize and split by sentences using the Stanford CoreNLP toolkit[20] [Manning et al. 2014]. The same selection procedure has been applied to the crawled texts to select additional 2,150 sentences with 20,000 words, which are then annotated according to the same adaption of Universal Dependencies on Singlish. This extended treebank enables more insightful syntactic analysis on Singlish-specific grammars as illustrated in section 3.4 and helps to further boost the performance for various Singlish POS tagger and dependency parser models as shown in sections 4 and 5. We name this extended dataset **STB-EXT**.

3.4 Case Studies on Singlish-specific Syntax

With the major extension of the Singlish treebank in both size and diversity in contents, more Singlish-specific syntactic constructions are observed and substantiated by increasing frequencies. A full list of extended imported vocabularies and Singlish expressions and their meanings are tabulated in Tables **??** to **??** in Appendix **??**. In this sections, we specifically discuss 6 unique Singlish syntactic constructions that are more commonly observed in the extended Singlish treebank and the adaption of Universal Dependencies on them:

- "**de**" and "**le**" are the Hanyu Pinyin Romanization(Abbreviated to: Pinyin) of common interjections in Chinese[21], which appears often in Singlish texts due its Chinese origin. Accordingly, they have the same POS tags, "*X*" or "*PART*" depending on their specific usage, and the same dependency relations as exemplified in the UD Chinese corpus.
- "**kenna**" is an unique word in Singlish that carries multiple senses "*to get*" or "*going to*". The former serves as an auxiliary verb and is followed by a verb. Thus, it is annotated with the "*AUX*" POS tag and usually depend on the verb with the "auxpass" dependency relation. The latter serves as a verb when followed by the nominal object, and thus it has the "*VERB*" POS tag and often is the root of the sentence or the clause.
- "**tio**" is another unique word in Singlish that also carries multiple senses summarized in three main ways. First, if it means "*to get*" or "*accomplish*", it is equivalent to "*kenna*" and is thus annotated in the same way. Second, if it means "*accurately choose*", it is treated as a verb. Lastly, in the fixed expression "*tio boh ?*", as mentioned before, it is analogous to "*isn't it ?*" in English. Thus, in this case, it is treated as an interjection word used as a discourse particle at the end of a sentence.
- "**no need**" is a commonly used phrase in Singlish, where the POS of the word "*need*" is hard to determine to be either "*VERB*" or "*NOUN*" since "*no*" in Singlish often interchangeable with "*not*", especially when used at the start of a sentence or clause where no proceeding context is present to offer necessary syntactic clues. Thus, we establish a rule by differentiating its usage in two cases. First, if "*no need*" is followed by a verb, we consider it as a case where the particle "*to*" is considered as being omitted, and the verb is the head of an adjective clause modifying "*need*" as a "*NOUN*". In this case, "*no*" has the POS tag of "*DET*". On the other hand, if "*no need*" is followed by a noun, then "*need*" is naturally considered as a verb and "*no*" has a POS tag of "*PART*", equivalent to "*not*" which is used more often in standard English.

[20] https://stanfordnlp.github.io/CoreNLP/
[21] "de": 的; "le": 了.

- "**not**" as a common English word is extensively used in Singlish not accompanied with any auxiliary or copular verb. Some examples are "*Not ...* ", "*If not, ...*", "*If not ... , ...*", and the most typical "*Confirm not ...*"[22] expression. According to the usages in English, we have summarized the rule of thumb as: if "*not*" is used with an auxiliary verb, it has a POS of "PART". Otherwise, when used independently, it has a POS of "*ADV*", with exceptions when used in "*If not, ...*" and "*or not*". In both cases, it modifies its head with a "*advcl*" dependency relation. The corresponding word in Singlish is "**bo**" or "**boh**" and they are annotated in the same way according to this rule.
- Adjectives and adverbs are often used with a **duplication** as a tone intensifier in Singlish. For example, in the sentence "*Aiya , we are comparing what are cheap cheap mah*", "*cheap*" is duplicated to by the speaker to emphasize on such expression. This is unique in Singlish and the closest dependency relation can be applied is "*mwe*" for multi-word expression. However, we propose that it can be a unique language-specific construction in Singlish and possibly named as "*mwe: dup*".

4 PART-OF-SPEECH TAGGING

As previously mentioned, since POS tags are important features for dependency parsing, we first obtain the automatically predicted POS tags. Specifically, we investigate two approaches (i.e. neural stacking and neural multi-task learning) which leverage both the English and Singlish POS tagging datasets. As shown in Figure 3, the bi-LSTM network with a CRF layer (bi-LSTM-CRF) is chosen as the base model, it has shown state-of-the-art performance by globally optimizing the tag sequence [Chen et al. 2016; Huang et al. 2015]. For the neural stacking approach, we train a POS tagger for UD-Eng using the base model. Based on this English POS tagging model, we train a POS tagger for Singlish using the feature-level neural stacking model of Chen et al. [2016]. Both the English and Singlish models consist of an input layer, a feature layer, and an output layer. For the neural multi-task structure shown in Figure 5, the input layer and feature layer are shared by both the English and Singlish models, and each model has distinct output layers. Same with the multi-task model of Chen et al. [2016], model parameters are trained by both tasks.

4.1 Base Bi-LSTM-CRF POS Tagger

Input Layer: Each token is represented as a vector by concatenating a word embedding from a lookup table with a weighted average of its character embeddings given by the attention model of Bahdanau et al. [2014]. Following Chen et al. [2016], the input layer produces a dense representation for the current input token by concatenating its word vector and the ones for its surrounding context tokens in a window of finite size.

Feature Layer: This layer employs a bi-LSTM network to encode the input into a sequence of hidden vectors that embody global contextual information. Following Chen et al. [2016], we adopt bi-LSTM with peephole connections [Graves and Schmidhuber 2005].

Output layer: As shown in Figure 3, this is a linear layer over the feature layer, followed by a CRF layer, to predict the POS tags for the input words by maximizing the conditional probability of the sequence of tags given input sentence. The hidden state vectors from the linear layer are called the emission vectors.

4.2 POS Tagger with Neural Stacking

We adopt the deep integration neural stacking structure presented in Chen et al. [2016]. As shown in Figure 4, the distributed vector representation for the target word at the input layer of the

[22]Equivalent to: *I'm sure that ... is/do not.*

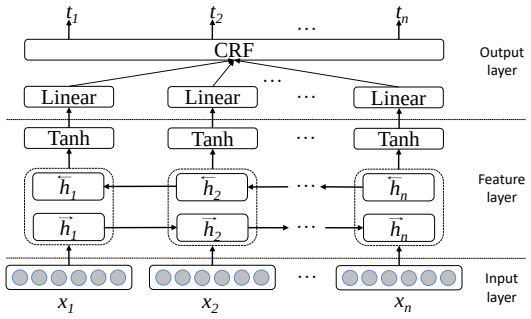

Fig. 3. Base POS tagger

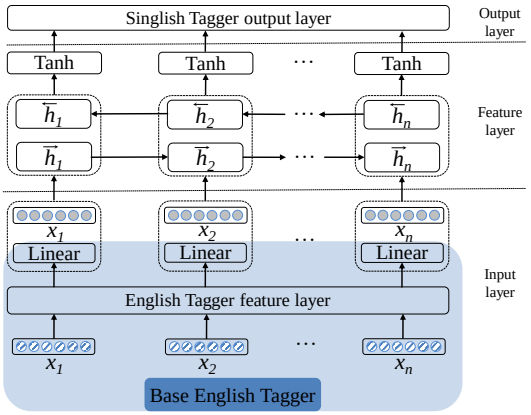

Fig. 4. Neural stacking POS tagger

Singlish Tagger is augmented by concatenating the emission vectors produced by the English Tagger with the original word and character-based embeddings, before applying the concatenation within a context window described in section 4.1. During training, the loss is back-propagated to all trainable parameters in both the Singlish Tagger and the pre-trained layers of the base English Tagger. At test time, the input sentence is fed to the whole integrated tagger model for inference.

4.3 POS Tagger with Neural Multi-task Learning

Following Chen et al. [2016], we investigate the multi-task learning for Singlish POS tagging. Figure 5 shows the multi-task structure, the English tagger and Singlish tagger share the same input layer and feature layer but use different output layers. In the training phase, sentences from both datasets are both fed into the same input layer and feature layer, but flow through the respective output layer of each task to compute the score of label sequences. The loss of each POS tagger is back-propagated to the output layer of the corresponding task and the shared feature layer and the input layer. The testing stage is similar to the base tagger but only the tagger output layer corresponding to the language of the input sentence is used.

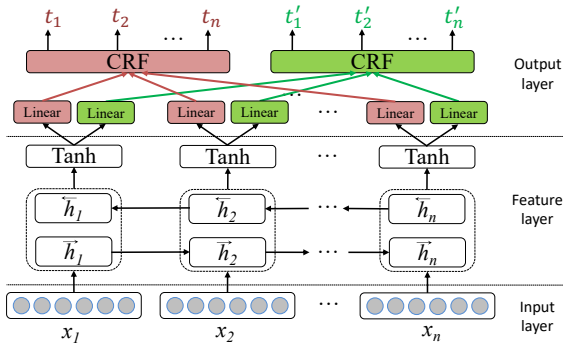

Fig. 5. Multi-task POS tagger

System	**STB-ACL**	**STB-EXT**
ENG-on-SIN	81.39	81.39
Base-ICE-SIN	78.35	88.27
Base-ICE-ENG+SIN	87.76	89.72
Stack-ICE-SIN	**89.50**	90.73
Multitask-ICE-SIN	87.83	**91.16**
Integrated-ICE-SIN	87.91	91.09

Table 2. POS tagging accuracies (%) for models trained on **STB-ACL** and **STB-EXT**

4.4 Results

Base English POS tagger: we followed Chen et al. [2016] and used the publicly available source code[23] to train a 1-layer bi-LSTM-CRF based POS tagger on UD-Eng, using 50-dimension pre-trained SENNA word embeddings [Collobert et al. 2011]. The hidden layer size was set to 300, the initial learning rate for Adagrad [Duchi et al. 2011] to 0.01, the regularization parameter λ to 10^{-6}, and the dropout rate to 15%. The tagger gives 94.84% accuracy on the UD-Eng test set after 24 epochs, chosen according to development tests, which is comparable to the state-of-the-art accuracy of 95.17% reported by Plank et al. [2016]. These settings were used to perform 10-fold jackknifing of POS tagging on the UD-Eng training set, with an average accuracy of 95.60%. This English POS tagger is used as the base English tagger in the neural stacking model described in Section 4.2.

Base Singlish POS tagger: Similarly, we trained a base Singlish POS tagger using the Singlish treebank alone with pre-trained word embeddings on the Singapore Component of the International Corpus of English (ICE-SIN) [Nihilani 1992; Ooi 1997], which consists of both spoken and written Singlish texts. Due to the limited amount of training data, the POS tagging accuracy on **STB-ACL** is significantly lower compared to all other models even with a larger dropout rate to avoid over-fitting during experiments, as shown by *Base-ICE-SIN* in Table 2[24]. However, the POS tagging accuracy is significantly improved from 78.35% to 88.27% using **STB-EXT** with 45.82% relative error reduction, which shows the effectiveness of the increased scale of training data.

[23] https://github.com/chenhongshen/NNHetSeq

[24] In order for fair comparison with the results trained on **STB-ACL** and also compare the relative amounts of improvement across different models, we restrict all our experiments to the same dev and test sets despite their relative small size. We leave for future explorations to perform significance test and experiments with the different dataset splits on **STB-EXT**.

Neural stacking: The neural stacking model has achieved the best accuracy of 89.50% using the **STB-ACL**[25], as is shown by *Stack-ICE-SIN* in Table 2, which corresponds to a 51.50% relative error reduction over its baseline Singlish model. Similarly, when training the neural stacking model on the **STB-EXT**, the POS tagging accuracy is further enhanced by 1.23% absolute accuracy improvement and 11.71% relative error reduction over its baseline Singlish POS tagger.

Multi-task model: The multi-task models can also improve the POS tagger on both **STB-ACL** and **STB-EXT** significantly, as shown by *Multitask-ICE-SIN* in Table 2. Comparing with the neural stacking model, the multi-task model has lower accuracy when trained on **STB-ACL** but outperforms the neural stacking model on **STB-EXT**.

This shows that the multi-task model benefits more from the extension of training data even with only one shared input layer compared to separate input layers in the neural stacking model. The reason may be that the **STB-EXT** dataset can provide sufficient samples for the multi-task network structure to extract more transferable features between English and Singlish which leads to a better performance. [26]

We also experimented with the integrated model by Chen et al. [2016] which utilizes both neural stacking and multi-task structures, as shown by *Integrated-ICE-SIN* in Table 2. However, it does not achieve further improvements, which is different from the observation by Chen et al. [2016]. Moreover, simply combining the UD-Eng and **STB-ACL** or **STB-EXT**datasets and training base POS tagger models yields improvements over the two baseline models train on a single language but still under-performs the stacking and multi-task models on both STBs. This again shows the availability of transferable knowledge from English to Singlish and also the better effectiveness of the transfer learning models.

In order to produce the automatically predicted POS tags for all train, development and test datasets, we use the best POS tagger, *Stack-ICE-SIN*, on **STB-ACL**, and *Multitask-ICE-SIN* on **STB-EXT** to perform 5-fold jackknifing on the training sets and label the development and test dataset respectively.

5 DEPENDENCY PARSING

We adopt the Dozat and Manning [2017] parser[27] as our base model, as displayed in Figure 6, and apply neural stacking and neural multi-task learning to achieve improvements over the baseline parser. Both the base and neural stacking models consist of an input layer, a feature layer, and an output layer. On the other hand, the multi-task model consist of shared input and feature layers but has distinct output layers for the English and Singlish parsers, respectively.

[25]We empirically find that using *ICE-SIN* embeddings in neural stacking model performs better than using English SENNA embeddings. Similar findings are found for the parser, of which more details are given in section 6.

[26]To further prove this consistency of improvement brought by dataset extension, we pre-trained another set of word embeddings using all the 420-million-word raw texts crawled from the Internet forums, called *Forum-SIN*. We used it with the the multi-task model and it leads to further improvements of Singlish POS tagging accuracy to 88.83% using the **STB-ACL** dataset and 91.45% using the **STB-EXT** dataset respectively. The *Forum-SIN* embeddings to a certain extent compensates the domain difference between the *ICE-SIN* embeddings and the Singlish dependency treebanks, which is more significant for the smaller **STB-ACL** dataset, as shown by the difference scale of improvement over *Multitask-ICE-SIN*: 1% for **STB-ACL** but 0.29% for **STB-EXT**. However, the *Forum-SIN* embeddings trained on forum texts are less concentrated on Singlish contents compared to *ICE-SIN* which was manually constructed with strict content control. On the other hand, the *Forum-SIN* is advantageous over the *ICE-SIN* embeddings by not only bridging the domain differences but also incorporating the context information of all texts annotated in the two treebanks during pre-training.

Therefore, to avoid the extra variety and preserve the fair comparison between only different models and datasets regarding knowledge transfer between language syntactic constructions, we make this set of embeddings available and leave further investigation on domain adaption to future work.

[27]https://github.com/tdozat/Parser

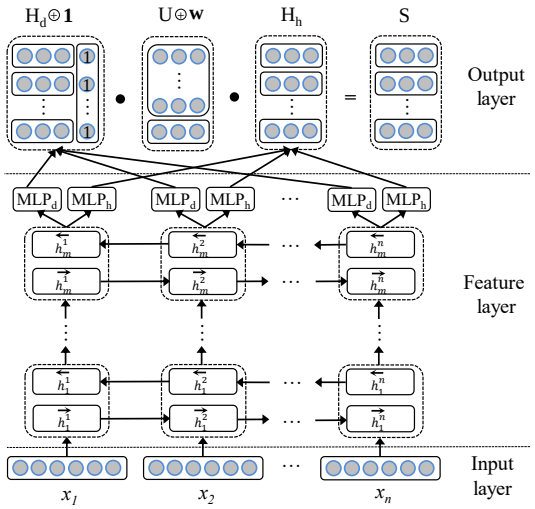

Fig. 6. Base parser model

5.1 Base Parser with Bi-affine Attentions

Input Layer: This layer encodes the current input word by concatenating a pre-trained word embedding with a trainable word embedding and POS tag embedding from the respective lookup tables.

Feature Layer: The two recurrent vectors produced by the multi-layer bi-LSTM network from each input vector are concatenated and mapped to multiple feature vectors in lower-dimension space by a set of parallel multilayer perceptron (MLP) layers. Following Dozat and Manning [2017], we adopt the Cif-LSTM cells by Greff et al. [2015].

Output Layer: This layer applies biaffine transformation on the feature vectors to calculate the score of the directed arcs between every pair of words. The inferred trees for input sentence are formed by choosing the head with the highest score for each word and a cross-entropy loss is calculated to update the model parameters.

5.2 Parser with Neural Stacking

Inspired by the idea of feature-level neural stacking [Chen et al. 2016; Zhang and Weiss 2016], we concatenate the pre-trained word embedding, trainable word and tag embeddings, with the two recurrent state vectors at the last bi-LSTM layer of the English Tagger as the input vector for each target word. In order to further preserve syntactic knowledge retained by the English Tagger, the feature vectors from its MLP layer is added to the ones produced by the Singlish Parser, as illustrated in Figure 7, and the scoring tensor of the Singlish Parser is initialized with the one from the trained English Tagger. The loss is back-propagated by reversely traversing all forward paths to all trainable parameter for training and the whole model is used collectively for inference.

5.3 Parser with Neural Multi-task Learning

Similar to the multi-task POS tagging structure, we also investigate the parser with neural multi-task joint training with English corpus. Figure 8 shows the structure of neural multi-task parser. However, since the multi-layer bi-LSTM network serves as the general feature extractor that captures the contextual information from for each input word by combining the hidden state vectors from both

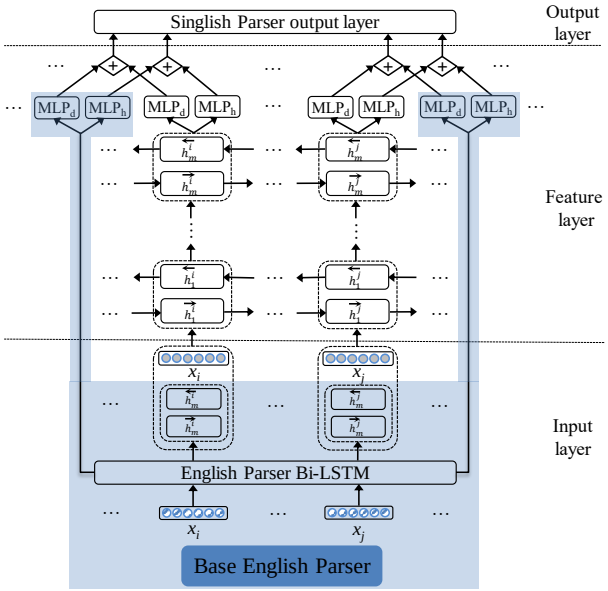

Fig. 7. Parser with neural stacking

directions through concatenation, the task-specific layers start from the set of MLP layers that maps the feature vectors in different lower-dimension spaces for different tasks (in this case English and Singlish parsers respectively). This is then followed by distinct output layers for constructing different parse trees for English and Singlish respectively. This network structure is different from the neural stacking model described in section 5.2 in two ways. The first is that the input layer and the feature layer is shared in the multi-task model instead of separate for neural stacking model which aims to capture the common syntactic features shared by both English and Singlish. Secondly, the model is jointly trained using both the English and Singlish treebanks instead of pre-training on the English base parser model and loaded to the base components in the neural stacking model. This aims to provide regularization to the input and feature layers that encode the transferable syntactic feature between English and Singlish from training on the English treebank while the Singlish-specific output layers are trained using the Singlish treebank.

6 EXPERIMENTS

6.1 Experimental Settings

We train a base English parser on UD-Eng with the default model settings in Dozat and Manning [2017]. It achieves a UAS of 88.83% and a LAS of 85.20%, which are close to the state-of-the-art 85.90% LAS on UD-Eng reported by Ammar et al. [2016], and the main difference is caused by us not using fine-grained POS tags. We apply the same settings for a baseline Singlish parser. We attempt to choose a better configuration of the number of bi-LSTM layers and the hidden dimension based on the development set performance, but the default settings turn out to perform the best. Thus we stick to all default hyper-parameters in Dozat and Manning [2017] for training the Singlish parsers.

We experimented with different word embeddings and further described in section 6.2. When using the neural stacking model, we fix the model configuration for the base English parser model and choose the size of the hidden vector and the number of bi-LSTM layers stacked on top based on the performance on the development set. It turns out that a 1-layer bi-LSTM with 900 hidden

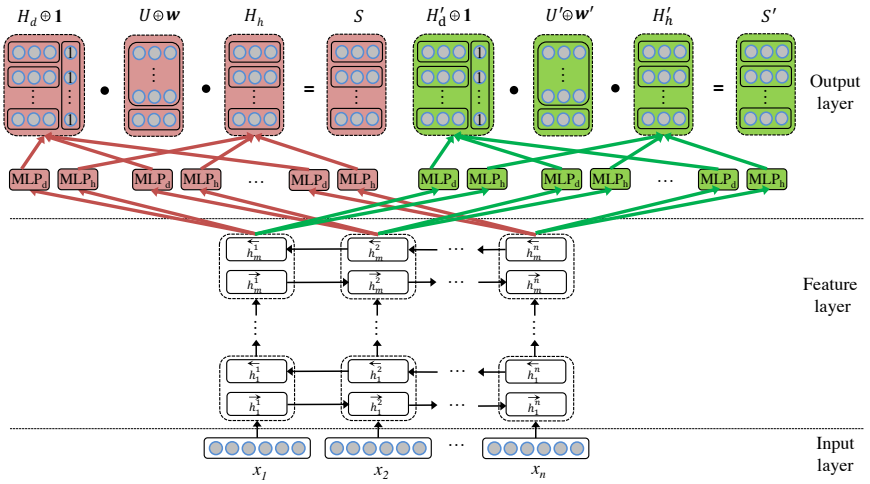

Fig. 8. Parser with multitask structure

	Sentences	Words	Vocabulary
GloVe6B	N.A.	6000m	400,000
Giga100M	57,000	1.26m	54,554
ICE-SIN	87,084	1.26m	40,532

Table 3. Comparison of the scale of sources for training word embeddings

dimension performs the best on the **STB-ACL** dataset and 2-layer bi-LSTM with 600 hidden dimension performs the best on the **STB-EXT** dataset. The bigger hidden layers accommodate the elongated input vector to the stacked bi-LSTM and the fewer number of recurrent layers helps to avoid over-fitting on the relatively small Singlish dependency treebanks compared to UD-Eng given the deep bi-LSTM English parser network at the bottom. The evaluation of the neural stacking and multi-task models is further described in below.

6.2 Investigating Distributed Lexical Characteristics

In order to learn characteristics of distributed lexical semantics for Singlish, we compare performances of the base Singlish dependency parser trained on **STB-ACL** using several sets of pre-trained word embeddings with the raw text sources summarized in Table 3: *GloVe6B*, large-scale English word embeddings[28], and *ICE-SIN*, Singlish word embeddings trained using GloVe [Pennington et al. 2014] on the ICE-SIN [Nihilani 1992; Ooi 1997] corpus. These two sets of embeddings capture the distributional semantics from English and Singlish respectively. However, due to the significant difference in the size of the corpus used for training the embeddings, in order for a fair comparison, we have trained another set of embeddings-*Giga100M*, a small-scale English word embeddings trained using GloVe [Pennington et al. 2014] with the same settings as the **ICE-SIN** embeddings on a comparable size of English data randomly selected from the English Gigaword Fifth Edition.

As shown in Table 4, the English Giga100M embeddings marginally improve the Singlish parser from the baseline without pre-trained embeddings and also using the UD-Eng parser directly on Singlish, represented as *ENG-on-SIN* in Table 4. With much more English lexical semantics

[28]Trained with Wikipedia 2014 the Gigaword. Downloadable from http://nlp.stanford.edu/data/glove.6B.zip

	Trained on English	
System	UAS	LAS
ENG-on-SIN	75.89	65.62

	Trained on Singlish			
	STB-ACL		**STB-EXT**	
System	UAS	LAS	UAS	LAS
Baseline	75.98	66.55	79.71	71.99
Base-Giga100M	77.67	67.23	-	-
Base-GloVe6B	78.18	68.51	83.02	75.21
Base-ICE-SIN	79.29	69.27	82.51	75.30

	Trained on Singlish and English			
	STB-ACL		**STB-EXT**	
System	UAS	LAS	UAS	LAS
ENG-plus-SIN	82.43	75.64	83.28	75.81
Stack-ICE-SIN	**84.47**	**77.76**	**85.57**	**79.12**
Multitask-ICE-SIN	79.12	70.46	82.60	76.06

Table 4. Dependency parser performances measured by UAS and LAS for models trained on English only, Singlish only, and both English and Singlish. Two Singlish training sets are used: **STB-ACL** and **STB-EXT**

being fed to the Singlish parser using the English *GloVe6B* embeddings, further enhancement is achieved. Nevertheless, the Singlish *ICE-SIN* embeddings lead to even more improvement, with 13.78% relative error reduction, compared with 7.04% using the English *Giga100M* embeddings and 9.16% using the English *GloVe6B* embeddings, despite the huge difference in sizes in the latter case.

This demonstrates the distributional differences between Singlish and English tokens, even though they share a large vocabulary.

Similar improvements has been observed for **STB-EXT**, as shown in Table 4, and both *ICE-SIN* and *GloVe6B* embeddings have lead to significant improvement over the baseline model. Furthermore, the Singlish *ICE-SIN* embeddings still yield comparable performance to the English *GloVe6B* embeddings with slightly lower UAS and slightly higher LAS on **STB-EXT**. This indicates that the distributional semantics from both the English and Singlish are beneficial for Singlish parsing.

The *Multi-task-SIN* yields lower performance on **STB-ACL** but comparable results with *ENG-plus-SIN* as shown in Table 4. With the additional output layer and corresponding parameters, the multi-task model confronts more limitations using the smaller **STB-ACL** but on the other hand is better at leveraging more training data to improve the performance[29].

6.3 Knowledge Transfer Using Neural Stacking

We train a parser with the neural stacking model and Singlish *ICE-SIN* embeddings on the **STB-ACL** dataset, which achieves the best performance among all the models, with a UAS of 84.47%, represented as *Stack-ICE-SIN* in Table 4, which corresponds to 25.01% relative error reduction compared to the comparable baseline, *Base-ICE-SIN*. This demonstrates that knowledge from English can be successfully incorporated to boost the Singlish parser. To further evaluate the effectiveness of the neural stacking model, we also trained a base model with the combination of UD-Eng and the Singlish treebank, represented as *ENG-plus-SIN* in Table 4, which is still outperformed by the neural stacking model.

[29]We tried with different corpus weighting ratio between UD-Eng and empirically found 1:1 works the best

System	UAS	LAS
Base-ICE-SIN	77.00	66.69
Stack-ICE-SIN	**82.43**	**73.96**

Table 5. Dependency parser performances by the 5-cross-fold validation

Besides, we performed a 5-cross-fold validation for the base parser with Singlish *ICE-SIN* embeddings and the parser using neural stacking, where half of the held-out fold is used as the development set. The average UAS and LAS across the 5 folds shown in Table 5 and the relative error reduction on average 23.61% suggest that the overall improvement from knowledge transfer using neural stacking remains consistent.

6.4 Improvements over Grammar Types

	Topic Prominence		Copula Deletion		NP Deletion		Discourse Particles		Others	
Percentage	8.65%		10.98%		12.14%		29.48%		38.73%	
	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS
ENG-on-SIN	78.15	62.96	66.91	56.83	72.57	64.00	70.00	59.00	78.92	68.47
Base-Giga100M	77.78	68.52	71.94	61.15	76.57	69.14	85.25	77.25	73.13	60.63
Base-ICE	81.48	72.22	74.82	63.31	**80.00**	73.71	85.25	77.75	75.56	64.37
Stack-ICE	**87.04**	**76.85**	**77.70**	**71.22**	**80.00**	**75.43**	**88.50**	**83.75**	**84.14**	**76.49**

Table 6. Error analysis with respect to grammar types

To analyze the sources of improvements for Singlish parsing using different model configurations, we conduct error analysis over 5 syntactic categories[30] using **STB-ACL**, including 4 types of grammars mentioned in section 3.2[31], and 1 for all other cases, including sentences containing imported vocabularies but expressed in basic English syntax. The number of sentences and the results in each group of the test set are shown in Table 6.

The neural stacking model leads to the biggest improvement over all categories except for a tie UAS performance on "*NP Deletion*" cases, which explains the significant overall improvement.

Comparing the base model with *ICE-SIN* embeddings with the base parser trained on UD-Eng, which contain syntactic and semantic knowledge in Singlish and English, respectively, the former outperforms the latter on all 4 types of Singlish grammars but not for the remaining samples. This suggests that the base English parser mainly contributes to analyzing basic English syntax, while the base Singlish parser models unique Singlish grammars better.

Similar trends are also observed on the base model using the English Giga100M embeddings, but the overall performances are not as good as using *ICE-SIN* embeddings, especially over basic English syntax where it undermines the performance to a greater extent. This suggests that only limited English distributed lexical semantic information can be integrated to help to model Singlish syntactic knowledge due to the differences in distributed lexical semantics.

6.5 Knowledge Transfer Using Neural Multi-task Learning

As is shown by *Multitask-ICE-SIN* in Table 4, *Multitask-ICE-SIN* has significantly lower accuracies compared with *ENG-plus-SIN* since **STB-ACL** is not large enough compared to UD_Eng for the

[30]The percentages add up to more than 100% since multiple labels are allowed for one sentence.
[31]The "Inversion" type of grammar is not analyzed since there is only 1 such sentence in the test set.

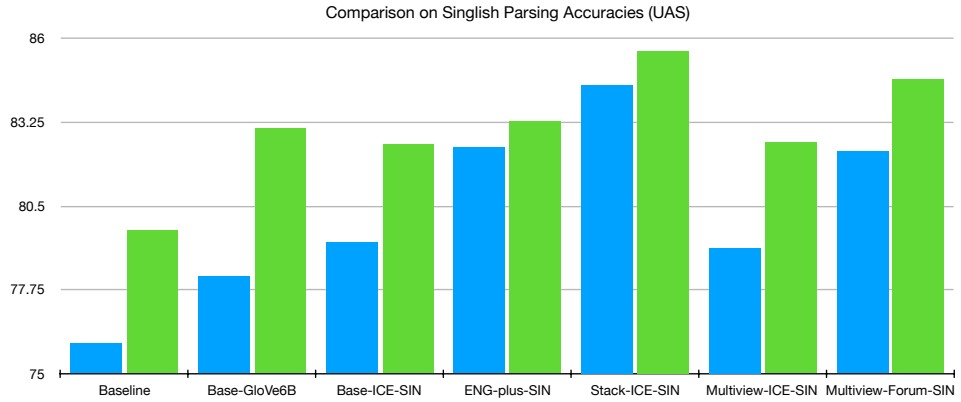

Fig. 9. Comparison on dependency parser performances between STB-ACL and STB-EXT

parser to extract sufficient transferable syntactic knowledge between Singlish and English by means of multi-task learning. This is substantially mitigated by the enlarged size of the Singlish training data when using **STB-EXT**, which leads to comparable accuracies with *ENG-plus-SIN*. This demonstrates the effectiveness of the extended dataset and the ability of the multi-task network structure on better knowledge transfer compared with the simple data mixing approach on a larger training dataset.

On the other hand, the Singlish parser with the multi-task model trained on both the two datasets has lower accuracies than the parser with the neural stacking structure (*Stack-ICE-SIN*). The main reason is that the separate input layer in the neural stacking model is better at capturing language-specific syntactic information compared to the shared input layer jointed trained by two languages. This defect can be partially compensated by the enlargement of the training dataset but the performance is still not satisfactory. Another hypothesis is that the domain difference between the Singlish *ICE-SIN* embeddings and the Singlish dependency treebanks is magnified when using a shared input layer.

6.6 Significance of Dataset Extension

Figure 9 has illustrated that all parsers trained on the **STB-EXT** dataset yield better results[32], which has proven the quality of the **STB-EXT** dataset. Besides, their consistency in the ranking of the parsing accuracies further substantiated our findings in our conference paper. Furthermore, we find that the improvement from training data extension is more significant when applying the multi-task model (LAS from 70.46% to 76.06% with 18.96% relative error reduction) than the neural stacking parser (LAS from 77.76% to 79.12% with 6.12% relative error reduction). This is consistent with our previous observation that the multi-task model benefits more from the extension of training data compared to the neural stacking model in POS tagging task.

By extending the size of the training data using **STB-EXT**, the neural stacking parser is also improved significantly with 17.50% relative error reduction compared to *Base-ICE-SIN* and with with 13.70% relative error reduction compared to *ENG-plus-SIN*, as is shown by *Stack-ICE-SIN* in Table 4. This is consistent with the POS tagging accuracy improvement described in Section 4 and has lead to the new SOTA on Singlish parsing accuracies with UAS 85.57 and LAS 79.12.

[32]Performance measured by LAS has the same trend as UAS and is excluded for simplicity

7 CONCLUSION

We have investigated dependency parsing for Singlish, an important English-based creole language, through annotations of a Singlish dependency treebank with 30,986 words in total and building an enhanced parser by leveraging on knowledge transferred from a 7-times-bigger English treebank of Universal Dependencies. We demonstrate the effectiveness of our dataset extension by improvements in experiment results using all base, neural stacking and multi-task models. Besides, we specifically explored multi-task models which boost the Singlish POS tagging accuracy and dependency parsing performance by joint learning transferable features between English and Singlish. We release the extended Singlish dependency treebank, **STB-EXT**, the trained model and the source code for the parser with free public access. Possible future work includes expanding the investigation to other regional languages such as Malay and Indonesian, and designing of better feature selection mechanisms to assist transfer learning between English and its low-resource creole languages.

ACKNOWLEDGMENTS

Yue Zhang is the corresponding author. We are very grateful for the detailed and constructive comments from all three reviewers, which are helpful to make this paper better.

REFERENCES

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah Smith. 2016. Many Languages, One Parser. *Transactions of the Association of Computational Linguistics* 4 (2016), 431–444. http://aclweb.org/anthology/Q16-1031

Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. 2016. Globally Normalized Transition-Based Neural Networks. In *Proceedings of the ACL 2016*. Association for Computational Linguistics, 2442–2452. https://doi.org/10.18653/v1/P16-1231

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv preprint* abs/1409.0473 (2014). http://arxiv.org/abs/1409.0473

Miguel Ballesteros, Chris Dyer, and A. Noah Smith. 2015. Improved Transition-based Parsing by Modeling Characters instead of Words with LSTMs. In *Proceedings of the EMNLP 2015*. Association for Computational Linguistics, 349–359. https://doi.org/10.18653/v1/D15-1041

Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. 2012. English Web Treebank LDC2012T13. (2012).

Danqi Chen and Christopher Manning. 2014. A Fast and Accurate Dependency Parser using Neural Networks. In *Proceedings of the EMNLP 2014*. Association for Computational Linguistics, 740–750. https://doi.org/10.3115/v1/D14-1082

Hongshen Chen, Yue Zhang, and Qun Liu. 2016. Neural Network for Heterogeneous Annotations. In *Proceedings of the EMNLP 2016*. Association for Computational Linguistics, 731–741. http://aclweb.org/anthology/D16-1070

Shay Cohen and A. Noah Smith. 2009. Shared Logistic Normal Distributions for Soft Parameter Tying in Unsupervised Grammar Induction. In *Proceedings of the NAACL-HLT 2009*. Association for Computational Linguistics, 74–82. http://aclweb.org/anthology/N09-1009

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12 (2011), 2493–2537. http://dl.acm.org/citation.cfm?id=2078186

Timothy Dozat and Christopher D. Manning. 2017. Deep Biaffine Attention for Neural Dependency Parsing, In International Conference on Learning Representations 2017. *arXiv preprint* abs/1611.01734. http://arxiv.org/abs/1611.01734

John C. Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research* 12 (2011), 2121–2159. http://dl.acm.org/citation.cfm?id=2021068

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015. A Neural Network Model for Low-Resource Universal Dependency Parsing. In *Proceedings of the EMNLP 2015*. Association for Computational Linguistics, 339–348. https://doi.org/10.18653/v1/D15-1040

Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and A. Noah Smith. 2015. Transition-Based Dependency Parsing with Stack Long Short-Term Memory. In *Proceedings of the ACL-IJCNLP 2015*. Association for Computational Linguistics, 334–343. https://doi.org/10.3115/v1/P15-1033

Kuzman Ganchev, Jennifer Gillenwater, and Ben Taskar. 2009. Dependency Grammar Induction via Bitext Projection Constraints. In *Proceedings of the ACL-IJCNLP 2009*. Association for Computational Linguistics, 369–377. http://aclweb.org/anthology/P09-1042

Douwe Gelling, Trevor Cohn, Phil Blunsom, and Joao Graca. 2012. The PASCAL Challenge on Grammar Induction. In *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*. Association for Computational Linguistics, 64–80. http://www.aclweb.org/anthology/W12-1909

Jennifer Gillenwater, Kuzman Ganchev, João Graça, Fernando Pereira, and Ben Taskar. 2010. Sparsity in Dependency Grammar Induction. In *Proceedings of the ACL 2010 (Short Papers)*. Association for Computational Linguistics, 194–199. http://www.aclweb.org/anthology/P10-2036

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18, 5 (2005), 602–610.

Klaus Greff, Rupesh Kumar Srivastava, Jan Koutník, Bas R. Steunebrink, and Jürgen Schmidhuber. 2015. LSTM: A Search Space Odyssey. *CoRR* abs/1503.04069 (2015). http://arxiv.org/abs/1503.04069

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual Dependency Parsing Based on Distributed Representations. In *Proceedings of the ACL-IJCNLP 2015*. Association for Computational Linguistics, 1234–1244. https://doi.org/10.3115/v1/P15-1119

Shinichi Harada. 2009. The roles of singapore standard english and singlish. *Information Research* 40 (2009), 70–82.

Kenneth Heafield, Ivan Pouzyrevsky, H. Jonathan Clark, and Philipp Koehn. 2013. Scalable Modified Kneser-Ney Language Model Estimation. In *Proceedings of the ACL 2013 (Short Papers)*. Association for Computational Linguistics, 690–696. http://aclweb.org/anthology/P13-2121

Sanjika Hewavitharana, Nguyen Bach, Qin Gao, Vamshi Ambati, and Stephan Vogel. 2011. *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Chapter CMU Haitian Creole-English Translation System for WMT 2011, 386–392. http://aclweb.org/anthology/W11-2146

Chang Hu, Philip Resnik, Yakov Kronrod, Vladimir Eidelman, Olivia Buzek, and B. Benjamin Bederson. 2011. *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Chapter The Value of

Monolingual Crowdsourcing in a Real-World Translation Scenario: Simulation using Haitian Creole Emergency SMS Messages, 399–404. http://aclweb.org/anthology/W11-2148

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv preprint* abs/1508.01991 (2015). http://arxiv.org/abs/1508.01991

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping Parsers via Syntactic Projection Across Parallel Texts. *Natural Language Engineering* 11, 3 (September 2005), 311–325. https://doi.org/10.1017/S1351324905003840

Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and Accurate Dependency Parsing Using Bidirectional LSTM Feature Representations. *Transactions of the Association of Computational Linguistics* 4 (2016), 313–327. http://aclweb.org/anthology/Q16-1023

Karen Lahousse and Béatrice Lamiroy. 2012. Word order in French, Spanish and Italian: A grammaticalization account. *Folia Linguistica* 46, 2 (2012), 387–415.

Jakob R. E. Leimgruber. 2009. *Modelling variation in Singapore English.* Ph.D. Dissertation. Oxford University.

Jakob R. E. Leimgruber. 2011. Singapore English. *Language and Linguistics Compass* 5, 1 (2011), 47–62. https://doi.org/10.1111/j.1749-818X.2010.00262.x

Jiwei Li, Thang Luong, Dan Jurafsky, and Eduard Hovy. 2015. When Are Tree Structures Necessary for Deep Learning of Representations?. In *Proceedings of the EMNLP 2015*. Association for Computational Linguistics, 2304–2314. https://doi.org/10.18653/v1/D15-1278

Lisa Lim. 2007. Mergers and acquisitions: on the ages and origins of Singapore English particles. *World Englishes* 26, 4 (2007), 446–473.

Yijia Liu, Yi Zhu, Wanxiang Che, Bing Qin, Nathan Schneider, and Noah A. Smith. 2018. Parsing Tweets into Universal Dependencies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*. 965–975. https://aclanthology.info/papers/N18-1088/n18-1088

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*. 55–60. http://www.aclweb.org/anthology/P/P14/P14-5010

Héctor Martínez Alonso, Željko Agić, Barbara Plank, and Anders Søgaard. 2017. Parsing Universal Dependencies without training. In *Proceedings of the EACL 2017*. Association for Computational Linguistics, 230–240. http://www.aclweb.org/anthology/E17-1022

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-Source Transfer of Delexicalized Dependency Parsers. In *Proceedings of the EMNLP 2011*. Association for Computational Linguistics, 62–72. http://aclweb.org/anthology/D11-1006

Ho Mian-Lian and John T. Platt. 1993. *Dynamics of a contact continuum: Singaporean English.* Oxford University Press, USA.

Makoto Miwa and Mohit Bansal. 2016. End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures. In *Proceedings of the ACL 2016*. Association for Computational Linguistics, 1105–1116. https://doi.org/10.18653/v1/P16-1105

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective Sharing for Multilingual Dependency Parsing. In *Proceedings of the ACL 2012*. Association for Computational Linguistics, 629–637. http://aclweb.org/anthology/P12-1066

Tahira Naseem, Harr Chen, Regina Barzilay, and Mark Johnson. 2010. Using Universal Linguistic Knowledge to Guide Grammar Induction. In *Proceedings of the EMNLP 2010*. Association for Computational Linguistics, Cambridge, MA, 1234–1244. http://www.aclweb.org/anthology/D10-1120

Paroo Nihilani. 1992. The international computerized corpus of English. *Words in a cultural context. Singapore: UniPress* (1992), 84–88.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the LREC 2016* (23-28). European Language Resources Association.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 Shared Task on Dependency Parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*. Association for Computational Linguistics, 915–932. http://www.aclweb.org/anthology/D/D07/D07-1096

Brendan T. O'Connor, Su Lin Blodgett, and Johnny Wei. 2018. Twitter Universal Dependency Parsing for African-American and Mainstream American English. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*. 1415–1425. https://aclanthology.info/papers/P18-1131/p18-1131

Vincent B Y Ooi. 1997. *Analysing the Singapore ICE corpus for lexicographic evidence.* ENGLISH LANGUAGE & LITERATURE. http://scholarbank.nus.edu.sg/handle/10635/133118

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the EMNLP 2014*. Association for Computational Linguistics, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss. In *Proceedings of the ACL 2016 (Short Papers)*. Association for Computational Linguistics, 412–418. https://doi.org/10.18653/v1/P16-2067

Manuela Sanguinetti, Cristina Bosco, Alberto Lavelli, Alessandro Mazzei, Oronzo Antonelli, and Fabio Tamburini. 2018. PoSTWITA-UD: an Italian Twitter Treebank in Universal Dependencies. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*.

Manuela Sanguinetti, Cristina Bosco, Alessandro Mazzei, Alberto Lavelli, and Fabio Tamburini. 2017. Annotating Italian Social Media Texts in Universal Dependencies. In *Proceedings of the Fourth International Conference on Dependency Linguistics, Depling 2017, Pisa, Italy, September 18-20, 2017*. 229–239. https://aclanthology.info/papers/W17-6526/w17-6526

Chun-Wei Seah, Hai Leong Chieu, Kian Ming Adam Chai, Loo-Nin Teow, and Lee Wei Yeong. 2015. Troll detection by domain-adapting sentiment analysis. In *18th International Conference on Information Fusion (Fusion) 2015*. IEEE, 792–799.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, D. Christopher Manning, Andrew Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the EMNLP 2013*. Association for Computational Linguistics, 1631–1642. http://aclweb.org/anthology/D13-1170

Anders Søgaard. 2012a. Two baselines for unsupervised dependency parsing. In *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*. Association for Computational Linguistics, 81–83. http://www.aclweb.org/anthology/W12-1910

Anders Søgaard. 2012b. Unsupervised dependency parsing without training. *Natural Language Engineering* 18, 2 (2012), 187–203. https://doi.org/10.1017/S1351324912000022

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual Word Clusters for Direct Transfer of Linguistic Structure. In *Proceedings of the NAACL-HLT 2012*. Association for Computational Linguistics, 477–487. http://aclweb.org/anthology/N12-1052

Hongmin Wang, Yue Zhang, GuangYong Leonard Chan, Jie Yang, and Hai Leong Chieu. 2017. Universal Dependencies Parsing for Colloquial Singaporean English. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 1732–1744. http://aclweb.org/anthology/P17-1159

David Weiss, Chris Alberti, Michael Collins, and Slav Petrov. 2015. Structured Training for Neural Network Transition-Based Parsing. In *Proceedings of the ACL-IJCNLP 2015*. Association for Computational Linguistics, 323–333. https://doi.org/10.3115/v1/P15-1032

Yuan Zhang and Regina Barzilay. 2015. Hierarchical Low-Rank Tensors for Multilingual Transfer Parsing. In *Proceedings of the EMNLP 2015*. Association for Computational Linguistics, 1857–1867. https://doi.org/10.18653/v1/D15-1213

Yuan Zhang and David Weiss. 2016. Stack-propagation: Improved Representation Learning for Syntax. In *Proceedings of the 54th ACL*. Association for Computational Linguistics, 1557–1566. https://doi.org/10.18653/v1/P16-1147

Hao Zhou, Yue Zhang, Shujian Huang, and Jiajun Chen. 2015. A Neural Probabilistic Structured-Prediction Model for Transition-Based Dependency Parsing. In *Proceedings of the ACL-IJCNLP 2015*. Association for Computational Linguistics, 1213–1222. https://doi.org/10.3115/v1/P15-1117

APPENDIX A AVAILABLE IN THE FULL PAPER AVAILABLE FROM TALLIP TRANSACTIONS