A Universal Phrase Tagset for Multilingual Treebanks

Aaron L.-F. Han[†‡], Derek F. Wong[†], Lidia S. Chao[†]
Yi Lu[†], Liangye He[†], Liang Tian[†]

[†] NLP[2]CT lab, Department of CIS, University of Macau
[‡] ILLC, University of Amsterdam

hanlifengaaron@gmail.com, {derekfw, lidiasc}@umac.mo
{takamachi660, wutianshui0515, tianliang0123}@gmail.com

Abstract. Many syntactic treebanks and parser toolkits are developed in the past twenty years, including dependency structure parsers and phrase structure parsers. For the phrase structure parsers, they usually utilize different phrase tagsets for different languages, which results in an inconvenience when conducting the multilingual research. This paper designs a refined universal phrase tagset that contains 9 commonly used phrase categories. Furthermore, the mapping covers 25 constituent treebanks and 21 languages. The experiments show that the universal phrase tagset can generally reduce the costs in the parsing models and even improve the parsing accuracy.

Keywords: Universal phrase tagset, Phrase tagset mapping, Multilingual treebanks, Parsing.

1 Introduction

In the past twenty years, many treebanks were developed, such as the Chinese treebank [1][2], English treebank [3][4], German treebank [5], French treebank [6], and Portuguese treebank [7][8], etc. There are mainly two types of parsing structure, dependency structure and phrase structure. For the phrase structure treebanks, to capture the characteristics of specific languages, they tend to design different phrase tagsets. The phrase categories span from ten to twenty or even more. This is indeed helpful in the syntax analysis of the special in-cased language. However, the different phrase tagsets also make inconvenience for the multilingual research. To facilitate the further research of multilingual tasks, this paper designs a refined universal phrase tagset using 9 common phrase categories. The mappings between the phrase tagsets from the existing phrase structure treebanks and the universal phrase tagset are conducted, which covers 25 treebanks and 21 languages.

To evaluate the designed universal phrase tagset and the phrase tagset mapping works, the parsing experiments are conducted for intrinsic analysis on the available corpora, including Penn Chinese treebank (CTB-7) from Linguistic Data Consortium

(LDC)[1] for Chinese, the Wall Street Journal (WSJ) Treebank from LDC for English, Floresta[2]-bosque Treebank for Portuguese, Euro-Fr[3] corpus for French, and Negra[4] Treebank [5] for German.

2 Proposed Universal Phrase Tagset

The universal phrase tagset is designed to include Noun phrase (NP), Verbal phrase (VP), Adjectival phrase (AJP), Adverbial phrase (AVP), Prepositional phrase (PP), sentence or sub-sentence (S), Conjunction phrase (CONJP), Coordinated phrase (COP), and others (X) for covering the list marker, interjection, URL, etc.

The refined phrase tagset mapped from 25 existing treebanks to the universal phrase categories is detailed in Table 1. Most of the mapping is easily understood except for some special cases. For instance, the Chinese phrase tag DNP (phrase formed by *something+associative 的*) is mapped into AJP because it specifies the adjective phrase. The Chinese phrase tag DVP (*something+DEV 地*) is mapped into AVP due to that the character "地" specifies an adverbial phrase in Chinese.

3 Parsing Experiments

To validate the effectiveness of the universal phrase tagset, we conduct the evaluation on the parsing task. We first construct the parsing models based on original treebanks, training and testing. Then, the experiment is repeated by replacing the treebanks with ones annotated with the universal phrasal tags. The parsing experiments are conducted on the treebanks covering Chinese (CN), English (EN), Portuguese (PT), French (FR), and German (DE).

The experiments are based on the Berkeley parser [9], which focuses on learning probabilistic context-free grammars (PCFGs) to assign a sequence of words the most likely parse tree, and introduces the hierarchical latent variable grammars to automatically learn a broad coverage grammar starting from a simple initial one. The generated grammar is refined by hierarchical splitting, merging and smoothing. The Berkeley parser generally gains the best testing result using the 6^{th} smoothed grammar [10]. For a broad analysis of the experiments, we tune the parameters to learn the refined grammar by 7 times of splitting, merging and smoothing except 8 times for French treebank. The experiments are conducted on a server with the configuration stated in Table 2. The evaluation criteria include the cost of training time (hours), size of the generated grammar (MB), and the parsing scores, i.e., Labeled Precision (LPre), Labeled Recall (LRec), the harmonic mean of precision and recall (F1), and exact match (Ex).

[1] https://www.ldc.upenn.edu/
[2] http://www.linguateca.pt/floresta
[3] http://www.statmt.org/europarl/
[4] http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/

Table 1. The Mappings between the Universal Phrase Tagset and that of the Existing Treebanks

Universal Phrase Tag	English PennTreebank I [3]	English PennTreebank II [4]	Chinese PennTreebank [11]	Portuguese Floresta Treebank [7]	FrenchTreebank [6]	Japanese Treebank Tüba-J/S [12]
NP	NP, WHNP	NP, NAC, NX, QP, WHNP	NP, CLP, QP, LCP, WHNP	Np	NP	NPper, NPloc, NPtmp, NP, NP.foc
VP	VP	VP	VP, VCD, VCP, VNV, VPT, VRD, VSB	Vp	VN, VP, VPpart, VPinf	VP.foc, VP, VPcnd, VPfin
AJP	ADJP	ADJP, WHADJP	ADJP, DP, DNP	ap, adjp	AP	AP.foc, AP, APcnd
AVP	ADVP, WHADVP	ADVP, WHAVP, PRT, WHADVP	ADVP, DVP	advp	AdP	ADVP.foc, ADVP
PP	PP, WHPP	PP, WHPP	PP	Pp	PP	PP, PP.foc, PPnom, PPgen, PPacc
S	S, SBAR, SBARQ, SINV, SQ	S, SBAR, SBARQ, SINV, SQ, PRN, FRAG, RRC	IP, CP, PRN, FRAG, INC	fcl, icl, acl, cu, x, sq	SENT, Ssub, Sint, Srel, S	S, SS
CONJP		CONJP				
COP					COORD	
X	X	X, INTJ, LST, UCP	LST, FLR, DFL, INTJ, URL, X, UCP			ITJ, GR, err
Universal Phrase Tag	Danish Arboretum Treebank [38]	German NegraTreebank [5]	Spanish UAM Treebank [13]	Hungarian Szeged Treebank [39]	Spanish Treebank [14]	Swedish Talbanken05 [15]
NP	Np	NP, CNP, MPN, NM	HOUR, NP, QP, SCORE, TITLE	NP, QP	NP, MPN, MTC	CNP, NP
VP	vp, acl	VP, CVP, VZ, CVZ	VP	VP, INF_, INF0	SVC	CVP, VP
AJP	Ajp	AP, AA, CAP, MTA	ADJP	ADJP	AP	AP, CAP,
AVP	Dvp	AVP, CAVP	ADVP, PRED-COMPL	ADVP, PA_, PA0	AVP	AVP, CAVP,
PP	pp	PP, CAC, CPP, CCP	PP	PP	PP, MTP	CPP, PP
S	fcl, icl	S, CS, CH, DL, PSEUDO	CL, S	S	S, INC	CS, S
CONJP	cp			C0		
COP		CO		CP	CS, CNP, CPP, CAP, CAVP, CAC, CCP, CO	CONJP, CXP
X	par	ISU, QL		FP, XP		NAC, XP

Table 1. Continued

Universal Phrase Tag	Arabic PENN Treebank [16]	Korean Penn Treebank [17][18]	Estonian Arborest Treebank [40]	Icelandic IcePaHC Treebank [19]	Italian ISST Treebank [20][21]	Portuguese Tycho Brahe Treebank [22]
NP	NP, NX, QP, WHNP	NP	AN>, <AN, NN>, <NN,	NP, QP, WNP	SN	NP, NP-ACC, NP-DAT, NP-GEN, NP-SBJ, IP-SMC, NP-LFD, NP-ADV, NP-VOC, NP-PRN
VP	VP	VP	VN>, <VN, INF_N>, <INF_N	VP	IBAR	VB, VB-P
AJP	ADJP, WHADJP	ADJP, DANP		ADJP, ADJP-SPR	SA	ADJP, ADJP-SPR
AVP	ADVP, WHADVP	ADVP, ADCP	AD>, <AD	ADVP, ADVP-DIR, ADVP-LOC, ADVP-TMP, RP	SAVV	ADVP, WADVP
PP	PP, WHPP			PP, WPP, PP-BY, PP-PRN	SP, SPD, SPDA	PP, PP-ACC, PP-SBJ, PP-LFD, PP-PRN, PP-LOC
S	S, SBAR, SBARQ, SQ	S			F, SV2, SV3, SV5, FAC, FS, FINT, F2	RRC, CP, CP-REL, IP-MAT, IP-INF, IP-SUB, CP-ADV, CP-THT
CONJP	CONJP, NAC			CONJP	CP, COMPC	CONJP
COP			PN>, <PN		FC, COORD	
X	PRN, PRT, FRAG, INTJ, X, UCP	INTJ, PRN, X, LST, XP	<P, P>, <Q, Q>	LATIN	FP, COMPT, COMPIN	
Universal Phrase Tag	Hindi-Urdu Treebank [23]	Catalan AnCora Treebank [24]; [25]	Swedish Treebank [41]	Vietnamese Treebank [26]	Thai CG Tree-bank [27]	Hebrew [28]
NP	NP, NP-P, NP-NST, SC-A, SC-P, NP-P-Pred	sn	NP	NP, WHNP, QP	np, num, spnum	NP-gn-(H)
VP	VP, VP-Pred, V'	gv	VP	VP		PREDP, VP, VP-MD, VP-INF
AJP	AP, AP-Pred	sa	AP	AP, WHAP		ADJP-gn-(H)
AVP	DegP	sadv, neg	AVP	RP, WHRP		ADVP
PP		sp	PP	PP, WHPP	pp	PP
S		S, S*, S.NF.C, S.NF.A, S.NF.P, S.F.C, S.F.AComp, S.F.AConc, S.F.Acons, S.F.Acond, S.F.R,	ROOT, S	S, SQ, SBAR	s, ws, root	FRAG, FRAGQ, S, SBAR, SQ
CONJP		conj.subord, coord				
COP	CCP, XP-CC					
X	CP	interjeccio, morfema.verbal, morf.pron	XP	XP, YP, MDP		INTJ, PRN

Table 2. The Hardware Configuration

Memory	144 GB
CPU	Intel (R) Xeon (R) E5649 @ 2.53GHz (6 Cores)
Operating System	Ubuntu 64-bit

$$LPre = \frac{|\#correct\ constituent\ in\ guessed\ tree|}{|\#total\ constituent\ in\ guessed\ tree|} \quad (1)$$

$$LRec = \frac{|\#correct\ constituent\ in\ guessed\ tree|}{|\#total\ constituent\ in\ correct\ tree|} \quad (2)$$

$$F_1 = \frac{2 \times LPre \times LRec}{LPre + LRec} \quad (3)$$

$$Ex = \frac{|\#complete\ correct\ guessed\ tree|}{|\#total\ guessed\ tree|} \quad (4)$$

3.1 Parsing of Chinese

For Chinese, we use the Penn Chinese Treebank (CTB-7) [1] [2]. We adopt the standard splitting criteria for the training and testing data. The training documents contain CTB-7 files 0-to-2082, the development documents contain files 2083-to-2242, and the testing documents are files 2243-to-2447.

Table 3. The Parsing Results and Evaluation Scores on CTB-7

Grammar	Training cost		Evaluation score			
	Grammar (MB)	Training time	Precision	Recall	F1	Ex
Uni-gr-1	1.03	1m+54s	70.17	64.21	67.06	11.56
Ori-gr-1	1.16	1m+43s	73.57	68.31	**70.84**	**13.31**
Uni-gr-2	1.33	9m	76.78	72.24	74.44	12.98
Ori-gr-2	1.59	8m+29s	78.46	73.33	**75.81**	**14.5**
Uni-gr-3	1.97	24m+33s	81.37	76.7	78.97	18.7
Ori-gr-3	2.58	24m+28s	81.71	77.35	**79.47**	**19.57**
Uni-gr-4	3.39	1h+12m+19s	83.48	79.28	81.33	20.91
Ori-gr-4	5.01	1h+21m+4s	84.08	80.48	**82.24**	**21.37**
Uni-gr-5	6.97	4h+32m+26s	85.12	81.7	**83.37**	**23.03**
Ori-gr-5	11.08	17h+54m+53s	85.06	81.71	83.35	22.75
Uni-gr-6	15.54	16h+26m+24s	85.55	82.78	**84.15**	**24.68**
Ori-gr-6	27.26	23h+48m+1s	85.18	82.48	83.81	24.64
Uni-gr-7	34.53	56h+15m+16s	85.58	83.24	**84.4**	**25.33**
Ori-gr-7	65.55	94h+47m+18s	84.99	83.01	83.99	24.73

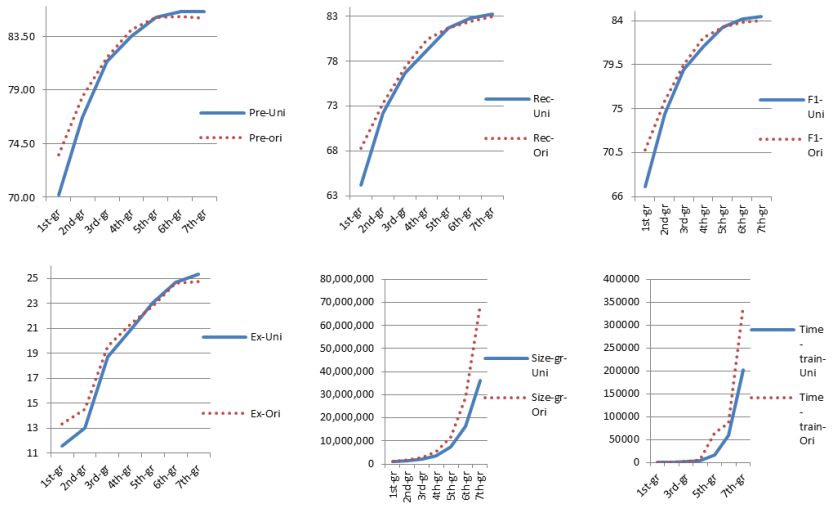

Fig. 1. Comparisons of Parsing Results of Chinese

We draw the corresponding learning curves in Figure 1 with byte and second for size and time. In Table 3, "Uni-gr-n" and "Ori-gr-n" mean the *n*th grammar using the universal and original tags respectively. The experiment shows that the testing scores of precision, recall and F1 are gradually higher using the refined grammars on universal phrase tagset than the scores using the original treebank tags, even though the beginning scores are lower from the first smoothed grammar. The exact match score using the universal phrase tagset also exceeds the corresponding score using original tagset after the 5th smoothed grammar. The highest precision, recall, F1 and exact scores are **85.58** (85.06), **83.24** (83.01), **84.4** (83.99), and **25.33** (24.73) respectively by using the universal phrase tagset (original phrase tags).

Furthermore, the training cost of the universal phrase tagset including the grammar size and the training time is much less than that of the original one especially for the latterly refined grammars. The grammar size (65.55 MB) and training time (94.79 hours) using the original tagset almost doubles that (**34.53 MB** & **56.25 hours**) of the universal one for the learning of 7th refined grammar.

3.2 Parsing of English

For English, we use the Wall Street Journal treebank from the LDC. The WSJ section 2-to-21 corpora are for training, section 22 for developing, and section 23 for testing [29]. The learning curves of EN for training cost and testing scores are shown in Figure 2. The experiment results on EN are similar to that on CN. The highest testing scores of precision, recall, and F1 are **91.45** (91.25), **91.19** (91.11) and **91.32** (91.18) respectively on universal phrase tags (and original phrase tagset). It takes **38.67** (51.64) hours and **30.72** (47.00) MB of memory during the training process for the 7th refined grammar respectively on universal (original) tagset.

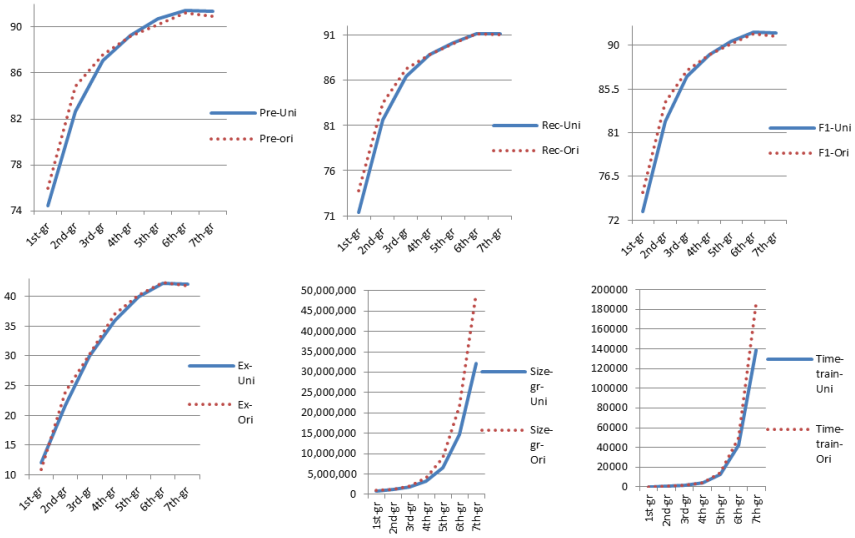

Fig. 2. Comparisons of Parsing Results of English

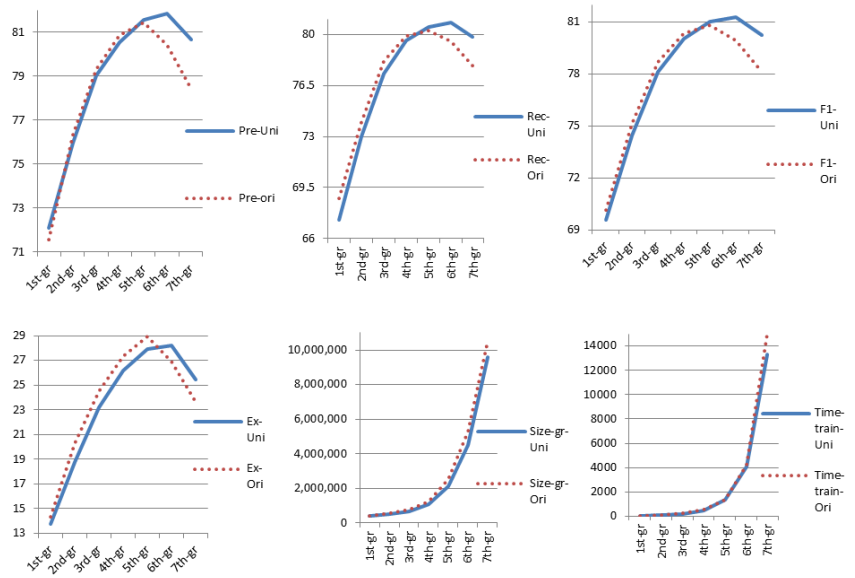

Fig. 3. Comparisons of Parsing Results of Portuguese

3.3 Parsing of Portuguese

The Bosque treebank is a subset of Floresta Virgem corpora [7] [8] with a size of 162,484 lexical units. We utilize 80 percent of the sentences for training, 10 percent

for developing and another 10 percent for testing. They are 7393, 939, and 957 sentences respectively. The learning curves of training cost and test scores are demonstrated in Figure 3.

The evaluation scores using the universal phrase tags are much higher than the ones using original tags after the 5th smoothed grammar. The highest scores of precision, recall and F1 are **81.84** (81.44), **80.81** (80.27) and **81.32** (80.85) respectively on universal (original) tags. It takes **3.69** (4.16) hours and **9.17** (10.02) MB of memory during the training process for the 7th refined grammar respectively on universal (original) tagset.

3.4 Parsing of German

We utilize the version 2.0 of Negra corpus [5] for German parsing, which consists of 355,096 tokens and 20,602 sentences German newspaper text with completely annotated syntactic structures. We use 80 percent (16,482 sentences) of corpus for training, 10 percent (2,060 sentences) for developing and 10 percent (2,060 sentences) for testing. The learning curves are shown in Figure 4.

The evaluation scores of DE language using universal phrase tags are slightly higher than the ones using original tags. The highest scores of precision, recall, F1 are **81.35** (81.23), **81.03** (81.02), and **81.19** (81.12) respectively on universal (original) tags. Different from the wining of synthetically F1 score, the exact matched sentence score on DE language is generally lower using the universal tags than using the original tags, and the generated grammar size becomes larger after 5th smoothing using the universal tags than the sizes using the original tagset.

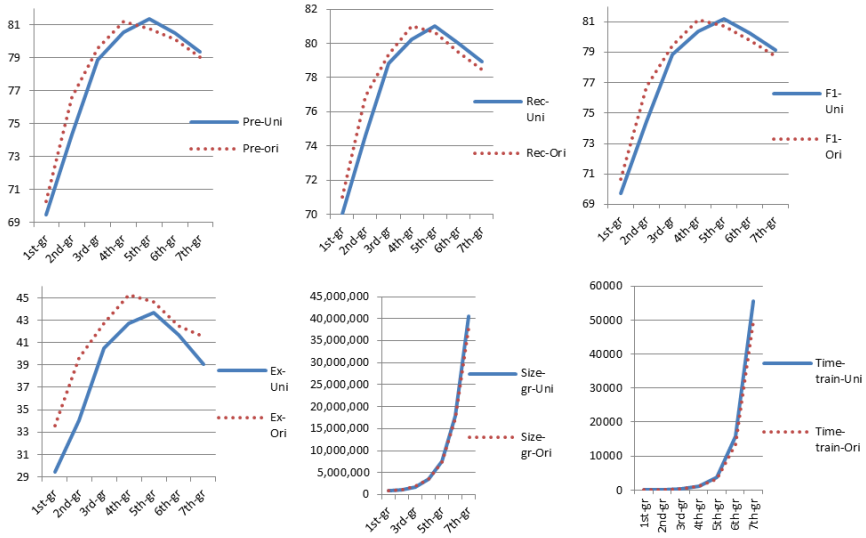

Fig. 4. Comparisons of Parsing Results of German

3.5 Parsing of French

Different with previous standard language treebanks, which are available by license agreement for research or commercial purpose, to generate a usable and reliable French treebank corpus, we first extract 20,000 French sentences from Europarl corpora that are from the proceedings of the European Parliament. Then, we parse the French plain text using the Berkeley French grammar "fra_sm5.gr" [10] with the parsing accuracy around 0.80. The parsed Euro-Fr corpus is used for the training stage.

For the developing and testing corpora, we use the WMT12 and WMT13 French plain text from the international workshop of statistical machine translation by SIGMT[5]. They contain 3,003 and 3,000 sentences respectively, which are parsed by the same parser. The experiment results of learning curves are shown in Figure 5. The evaluation results of FR show that the comprehensive F1 score using the universal tagset can also finally win the one using the original tagset, even though the exact match score is lower as the DE language. Similarly, the training cost using the universal tagset is much less. The highest precision, recall, and F1 scores are **80.49** (80.34), 80.93 (**80.96**), and **80.71** (80.64) respectively using universal (original) tagset.

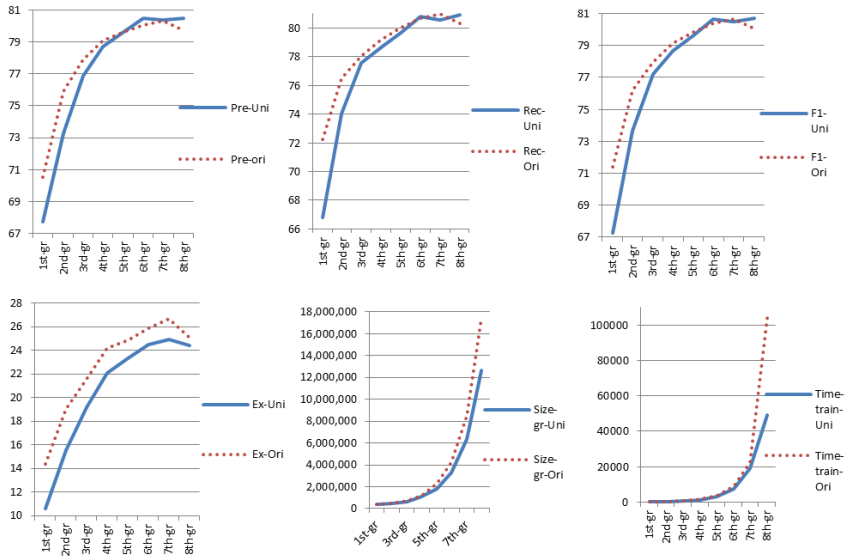

Fig. 5. Comparisons of Parsing Results of French

It takes **13.66** (28.91) hours and **12.07** (16.66) MB of memory during the training process for the 8th refined grammar respectively on universal (original) tagset.

[5] http://www.sigmt.org/

4 Related Work

Han et al. [30] proposed a universal phrase tagset and designed the mapping between the universal tagset and the ones of French and English Treebank. However, we extended the tagset mapping into 25 existing treebanks coving 21 languages; furthermore, we evaluated the effectiveness of the designed tagset mapping by parsing experiments on five available treebanks in this work. Other related work about phrase structures include [31], [32], and [33]. Naseem et al. [36] employed some manually specified universal dependency rules for grammar induction and achieved improvement in dependency parsing. McDonald et al. [37] designed a universal annotation approach for dependency treebanks. Rambow et al. [34] conducted a research about parallel syntactic annotation for multiple languages. Petrov et al. [35] developed a universal part-of-speech (PoS) tagset containing 12 commonly used PoS tags.

5 Conclusion

To facilitate the future researches in multilingual tasks, we have designed a refined universal phrase tagset and the tagset mapping from existing 25 treebanks into the universal tagset. To validate the designed work, evaluations are performed on parsing experiments. The evaluation on a range of language treebanks shows that the universal phrase tagset can generally improve the highest precision, recall and F1 testing score, especially on the Portuguese language, and reduce the training time and the size of generated grammar, especially on the Chinese, English and French languages. In the future, we plan to evaluate the parsing on more language treebanks, and utilize the universal phrase tagset into other multilingual applications.

Acknowledgments. The authors thank the anonymous reviewers for helpful comments. The work is partially supported by the Science and Technology Development Fund of Macau and the Research Committee of the University of Macau, the Reference No. MYRG076 (Y1-L2)-FST13-WF, and MYRG070 (Y1-L2)-FST12-CS.

References

1. Fei Xia, Martha Palmer, Nianwen Xue, et al. 2000. Developing Guidelines and Ensuring Consistency for Chinese Text Annotation, *Proceedings of LREC.*
2. Nianwen Xue, Fei Xia, Fu-Dong Chiou, and Martha Palmer. 2005: The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus. *Natural Language Engineering*, 11(2), 207-238.
3. Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: the penn treebank. *Comput. Linguist.* 19(2), 313-330.
4. Ann Bies, Mark Ferguson, Karen Katz and Robert MacIntyre. 1995. Bracketing Guidelines for Treebank II Style Penn Treebank Project. Technical paper.

5. W. Skut, B. Krenn, T. Brants, and H. Uszkoreit. 1997. An annotation scheme for free word order languages. In *Conference on ANLP*.

6. A. Abeillé, L. Clément, and F. Toussenel. 2003. Building a Treebank for French. *Building and Using Parsed Corpora*. Kluwer Academic Publishers.

7. Susana Afonso, Eckhard Bick, Renato Haber, and Diana Santos. 2002. Floresta sintá(c)tica: a treebank for Portuguese. In *Proceedings of LREC 2002*, pp.1698-1703.

8. Cláudia Freitas, Paulo Rocha and Eckhard Bick. 2008. Floresta Sintá(c)tica: Bigger, Thicker and Easier. In *Computational Processing of the Portuguese Language Conference*.

9. Slav Petrov and Dan Klein. 2007. Improved Inference for Unlexicalized Parsing. *NAACL*.

10. Slav Petrov. 2009. Coarse-to-Fine Natural Language Processing. PHD thesis.

11. Nianwen Xue, Zixin Jiang. 2010. Addendum to the Chinese Treebank Bracketing Guidelines(CTB7.0). Technical papar. University of Pennsylvania.

12. Yasuhiro Kawata, and Julia Bartels. 2000. Stylebook for the Japanese Treebank in VERBMOBIL. University Tubingen, Report 240.

13. Antonio Moreno, Susana López, Manuel Alcántara. 1999. Spanish Tree Bank: Specifications, Version 5. Technical paper.

14. M. Volk. 2009. Spanish Expansion of a Parallel Treebank. Technical paper.

15. Joakim Nivre, Jens Nilsson and Johan Hall. 2006. Talbanken05: A Swedish Treebank with Phrase Structure and Dependency Annotation. In *Proceedings of LREC*.

16. Ann Bies and Mohamed Maamouri. 2003. Penn Arabic Treebank Guidelines. Technical report.

17. Chung-hye Han, Na-Rae Han, Eon-Suk Ko. 2001. Bracketing Guidelines for Penn Korean TreeBank. Technical Report, IRCS-01-10.

18. Chung-hye Han, Na-Rae Han, Eon-Suk Ko, Heejong Yi and Martha Palmer. 2002. Penn Korean Treebank: Development and Evaluation, *Proceedings of PACLIC*.

19. Joel C. Wallenberg, Anton Karl Ingason, Einar Freyr Sigurðsson and Eiríkur Rögnvaldsson. 2011. Icelandic Parsed Historical Corpus (IcePaHC). Version 0.9. technical report.

20. S. Montemagni, Barsotti F., Battista M., et al. 2000. The Italian Syntactic-Semantic Treebank: Architecture, Annotation, Tools and Evaluation. *Proceedings of the COLING Workshop on Linguistically Interpreted Corpora*. pp. 18–27.

21. Simonetta Montemagni, Francesco Barsotti, Marco Battista, et. al. 2003. Building the Italian Syntactic-Semantic Treebank. in Anne Abeillé (ed.), *Building and using Parsed Corpora, Language and Speech series*, Kluwer, Dordrecht, chapter 11. pp. 189-210.

22. Charlotte Galves, and Pablo Faria. 2010. Tycho Brahe Parsed Corpus of Historical Portuguese. Technical.

23. Rajesh Bhatt, with Annahita Farudi and Owen Rambow. 2012. Hindi-Urdu Phrase Structure Annotation Guidelines. Technical Paper.

24. Montserrat Civit, Ma Antònia Martí 2004. Building cast3lb: A Spanish treebank. *Research on Language & Computation*, 2(4):549–574. Springer.

25. Taulé M., Martí M.A , Recasens, M. (2008). AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of LREC* 2008. Marrakech, Morocco.

26. Phuong-Thai Nguyen, Xuan-Luong Vu, et al. 2009. Building a large syntactically-annotated corpus of Vietnamese. *Lingu.Annotation Workshop*, 182-185.

27. T. Ruangrajitpakorn, K. Trakultaweekoon, and T. Supnithi. A syntactic resource for thai. 2009. CG treebank. In *Workshop on Asian Language Resources*, page 96–101, 2009.

28. Khalil Sima'an, Alon Itai, Yoad Winter, Alon Altman and Noa Nativ. 2001. Building a tree-bank of modern Hebrew text. *Journal Traitement Automatique des Langues*. *Special Issue on Natural Language Processing and Corpus Linguistics* 42(2), 347-380.

29. Slav Petrov, Leon Barrett, Romain Thibaux, Dan Klein. 2006. Learning Accurate, Compact, and Interpretable Tree Annotation. *COLING and 44th ACL*, pages 433–440.

30. Aaron L.-F. Han, Derek F. Wong, Lidia S. Chao, Liangye He, Shuo Li and Ling Zhu. 2013. Phrase Tagset Mapping for French and English Treebanks and Its Application in Machine Translation Evaluation. *LNCS Vol. 8105*, pp 119-131.

31. Robert D. Van Valin and Randy J. Lapolla. 2002. *Syntax, Structure, Meaning and Function*. Cambridge University Press.

32. Andrew Carnie. 2002. Syntax: A Generative Introduction (Introducing Linguistics). Blackwell Publishing.

33. Frederick J. Newmeyer. 2005. Possible and Probable Languages: A Generative Perspective on Linguistic Typology. Oxford University Press.

34. Owen Rambow, Bonnie Dorr, David Farwell, et al. 2006. Parallel syntactic annotation of multiple languages. In *Proceedings of LREC*.

35. Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A Universal Part-of-Speech Tagset, *Proceedings of the Eight LREC*.

36. Tahira Naseem, Harr Chen, Regina Barzilay, and Mark Johnson. 2010. Using universal linguistic knowledge to guide grammar induction. In *Proc. of EMNLP*.

37. Ryan McDonald; Joakim Nivre; Yvonne Quirmbach-Brundage; et al. 2013. Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of ACL*.

38. Danish Arboretum corpus. Arboretum: A syntactic tree corpus of Danish. Accessed December 2013. http://corp.hum.sdu.dk/arboretum.html

39. Hungarian Szeged Treebank. Szeged Treebank 2.0: A Hungarian natural language database with detailed syntactic analysis. Hungarian linguistics at the University of Szeged.

40. Eckhard Bick, Heli Uibo, and Kadri Muischnek. Preliminary experiments for a CG-based syntactic tree corpus of Estonian. Accessed Dec. 2013. http://corp.hum.sdu.dk/tgrepeye_est.html

41. Swedish Treebank Syntactic Annotation. Swedish Treebank. Online project. http://stp.lingfil.uu.se/~nivre/swedish_treebank/ accessed March, 2014.