# HuaYu: A Word-Segmented and Part-of-Speech Tagged Chinese Corpus<sup>1</sup>

# Sun Maosong\*, Sun Honglin<sup>†</sup>, Huang Changning\*, Zhang, Pu<sup>†</sup> Xing Hongbing<sup>†</sup>, Zhou Qiang\*

\*The State Key Laboratory of Intelligent Technology and Systems Tsinghua University, Beijing 100084, P. R. China sms@s1000e.cs.tsinghua.edu.cn

<sup>†</sup>Language Information Processing Institute Beijing Language and Culture University, Beijing 100084, P. R. China zhangpu@blcu.edu.cn

#### Abstract

As the outcome of a 3-year joint effort of Department of Computer Science, Tsinghua University and Language Information Processing Institute, Beijing Language and Culture University, Beijing, China, a word-segmented and part-of-speech tagged Chinese corpus with size of 2 million Chinese characters, named HuaYu, has been established. This paper firstly introduces some basics about HuaYu in brief, as its genre distribution, fundamental considerations in designing it, word segmentation and part-of-speech tagging standards. Then the complete list of tag set used in HuaYu is given, along with typical examples for each tag accordingly. Several pieces of annotated texts in each genre are also included at last for reader's reference.

# 1. Introduction

Corpus with part-of-speech annotation plays very important role in many natural language processing tasks, for instances, part-of-speech-tagging, compound noun analysis, chunking, and collocation extraction. Annotated corpora of English texts, e.g., LOB corpus, Brown corpus, have been intensively utilized by researchers of the above fields in last decade. On contrary, a widely accepted (or, 'standard') large scale annotated Chinese corpus is still not available to date. Chen et al (1996) established a wellknown Chinese corpus, SINACA corpus, with more than 3.5 million part-of-speech tagged tokens, however, the texts collected in this corpus are with Taiwan styles. which only represent language phenomena in part area of China. From the mid of 90's, much effort on constructing annotated Chinese corpora has been continuously paid in mainland of China. There are mainly two separate research teams working on this domain, one is from the Institute of Computational Linguistics of Peking University, another one is ours, from the State Key Laboratory of Intelligent Technology and Systems of Tsinghua University, and the Language Information Processing Institute of Beijing Language and Culture University. The work of these two teams resulted in two annotated corpora: Peking University corpus and HuaYu corpus. HuaYu corpus, built up by the latter team, will be addressed here. Readers who are interested in Peking University corpus please refer to the paper of Hu&Yu, 2000).

# 2. Basics of HuaYu

### 2.1. Distribution of *HuaYu*

HuaYu takes full article as basic unit of text collections. Articles in HuaYu are mainly published in 90's, and are distributed as four genres, i.e., literature,

news, academy and practical writing. firstly, we used CSeg&Tag1.0, a word segmentation and part-of-speech tagging integrated system developed by Tsinghua University (Sun *et al*, 1997), to automatically process the raw corpus. The output of CSeg&Tag1.0 was then proofread by human annotators. The distribution of HuaYu is summarized in table 1 (raw corpus) and table 2 (annotated corpus) respectively.

Genre	# of articles	# of Chinese	Ratio
		characters	
Literature	295	880,057	44%
News	376	600,490	30%
Academy	29	402,623	20%
Practical	258	119,488	6%
writing			
Total	958	2,002,658	100%

Table: 1 Distribution of the raw corpus

Genre	# of	# of words	Ratio
	punctuation		
Literature	148,453	760,337	48%
News	86,163	438,095	28%
Academy	52,823	278,728	18%
Practical	28,727	91,929	6%
writing			
Total	316,116	1,569,089	100%

Table 2: Statistics of HuaYu

#### 2.2. Aim of HuaYu

The aim of HuaYu is to be as either training set or testing set in particular for Chinese word segmentation and part-of-speech tagging algorithms. For example, all

<sup>&</sup>lt;sup>1</sup> This work is supported by National Natural Science Foundation of China and National Basic Research Development Scheme(973) of China.

character strings with ambiguous segmentations and their underlying pattern can be easily found from HuaYu, then we'll able to make use of machine learning techniques to extract relevant knowledge out of the corpus, so as to improve the performance of the existing models. Another example is, we further divide person names in Chinese texts into many subcategories: Chinese full name (including Chinese-style full name of Korean, Vietnamese etc), Chinese given name (including Chinese-style given name of Korean, Vietnamese etc), Chinese surname (including Chinese style surname of Korean, Vietnamese etc), Chinese name of minority nationalities, transliterated foreign name as well as Japanese name (including full name, surname and given name). As indicated by (Sun et al, 1997), person name identification will cause additional segmentation ambiguities, i.e. ambiguities among person name candidates and common words as well as ambiguities among different person name candidates (different from ambiguities among common words). The relevant detailed information in HuaYu will be definitely helpful for working out a robust person name identification module in word segmentation programs.

Another aim of HuaYu is to serve as data preparation for a ongoing research project of Tsinghua university to construct a Chinese tree bank with 10,000 representative sentences (Zhou & Sun, 1999). Obviously, a 'good' tree bank should cover variety of grammatical phenomena as much as possible in terms of their distributions in real texts. Because of the critical role of verb in grammatical analysis, plenty of verb sub-categorization information, as shown in table 3, has been incorporated into HuaYu. With this kind of information, it is very easy to write a computer program to randomly select sentences from HuaYu with the expected percentage of sentence patterns (Note, the expected percentage is estimated via HuaYutoo).

SubCat.	POS	Descriptions
1	Vgi	Intransitive verb
2	Vgn	verb taking an NP as object
3	Vgv	verb taking an VP as object
4	Vga	verb taking an AP as object
5	Vgs	verb taking an embedded sentence
		as object
6	Vgd	verb taking double objects
7	Vgj	Verb taking a pivotal sentence
8		

Table 3: The common verb subcategories

#### 2.3. Word segmentation standard and part-ofspeech tagging standard

The boundary between word and phrase in Chinese is quite fuzzy. No any linguistic rule proposed by linguists can work in a deterministic and systematic way if it is applied to language of real world. Consequently, the largest obstacle encountered in manually segmenting HuaYu is how to make segmentation consistent over the entire corpus. We set up a word segmentation standard, with emphasis on consistency issue, as guidance of manual segmentation. We also explored quantitative analysis means such as mutual information and difference of t-test values (Sun *et al*, 1998) supported by a raw corpus of 250 million Chinese characters, as an absolutely necessary supplement of the qualitative linguistic analysis. The word segmentation standard of Chinese texts is too language-specific, thus we do not intend to extend it here.

We also designed a part-of-speech tagging standard for annotating unrestricted Chinese texts. Any type of possible tokens in texts, no matter whether it is of word, morpheme, Chinese character, or non-Chinese-character string, should be able to be assigned a tag without exception, as required by language computation in open environment. The complete list of the tag set involved in the part-of-speech tagging standard, will be given in section 3.

## 3. Tag set of *HuaYu*

Totally 115 tags are defined in the tag set of *HuaYu*:

- 1 noun (11 tags)
- 1.1. proper noun
- 1.1.1. person name
- 1.1.1.1. <u>npc</u>: Chinese full name(including Chinese-style full name of Korean, Vietnamese etc)
- 1.1.1.2. <u>npm</u>: Chinese given name(including Chinesestyle given name of Korean, Vietnamese etc) e.g. ,
- 1.1.1.3. <u>npx</u>: Chinese surname(including Chinese style surname of Korean, Vietnamese etc) e.g. , , , , ,
- 1.1.1.4. <u>npz</u>: Chinese name of minority nationalities
- e.g. , , , , 1.1.1.5. <u>npf</u>: transliterated foreign name e.g. , , , , ,
- 1.1.1.6. <u>npj</u>: Japanese name(including full name, surname and given name)
  - e.g. , , ,
- 1.1.2. npu: organization
- e.g. , , , , , , , , , , , , , , , 1.1.3. <u>nps</u>: geographical place e.g. , , , (), , , (), (), (), (),
- 1.1.4. <u>npr</u>: proper noun, others
- e.g. , , ( )
- 1.2. <u>ng</u>: common noun
- e.g.,,,, 1.3. <u>nl</u>: separated noun() e.g.\_,,\_,\_
- 2. <u>t</u>: time word (1 tag)

e.g. , , , , , , , ,

3. s: place word (1 tag)

e.g. , , , , ,

NOTE

- Words as ", , , , , " should be uniquely classified into the category of noun.
- 4. <u>f</u>: position word (1 tag)

e.g.	,	,	,	,	,	,	,	,	,	,	,	,	,	,	,	,	,	,	,
		,	,		,	,	,	,		,	,		,	,		,	,	,	,

5. verb (16 tags) 5.1. va: auxiliary verb e.g. , , , , , , , , , , , , , , 5.2. <u>vi</u>: link verb e.g. , , , , , , , , 5.3. <u>vf</u>: formal verb e.g. , , 5.4. vv: pseudo verb ", " used before VP, but not for direction e.g. 5.5. vl: separated verb, in pair with nl \_(nl) e.g. \_(vl) 5.6. general verb 5.6.1. vgp: verb used as the immediate modifier of an NP e.g. , \_\_\_\_ 5.6.2. vgx: verb used as the head of an NP, without ""structure as modifier e.g. 5.6.3. <u>vgh</u>: verb used as the head of an NP, with" "structure as modifier e.g. \_\_\_, 5.6.4. vgc: verb used as complement e.g. , \_\_\_\_, \_\_\_\_, 5.6.6. transitive verb 5.6.6.1. vgn: verb taking an NP as object e.g. , \_\_\_\_ , \_\_\_\_ 5.6.6.2. vgv: verb taking an VP as object e.g. \_\_\_\_ , \_\_\_ 5.6.6.3. vga: verb taking an AP as object e.g. , \_\_\_ , \_\_\_ , \_\_ , 5.6.6.4. vgs: verb taking an embedded sentence as object e.g.\_\_\_\_\_, \_\_\_\_ 5.6.6.5. vgd: verb taking double objects e.g.:\_\_\_\_\_\_,\_\_\_ , 5.6.6.6. <u>vgj</u> verb taking a pivotal sentence e.g.\_\_\_\_,\_\_\_\_, \_ ,\_ NOTE In all cases under general verb, sub-category of the verb is determined by its grammatical function in sentences. The difference between vgv and vgs is, the object part of the latter has a subject whereas that of the former hasn't. 6. adjective (2 tags) 6.1. az: adjective used as immediate adverbial e.g. , , ,

6.2. <u>a</u>: general adjective

e.g. , , , , , , , , , , , , , ,

NOTE

If an adjective takes any object, then it should be classified as verb.

7. z: state word (1 tag)

8. b: distinction word (1 tag) e.g. , , , , , , , , , , , , , , , , NOTE Some of distinction words may function as adverbial, in this case, they are regarded as adverbial, e.g., " should be regarded as distinction word in phase " "whereas as adverbial in phrase " ". 9. numeral (8 tags) 9.1. <u>mx</u>: digit e.g. , , , , , , , , , , , , , , , 9.2. mw: digit place e.g. , , , , , 9.3. mg: approximation numeral e.g. , , , , , 9.4. mm: numeral-measure word e.g. , , , 9.5. ms: sequence word, " "" " and their combinations e.g. , 9.7. mam: infix of numeral e.g. \_\_\_\_, \_\_\_, \_\_\_ 9.8. mab: suffix of numeral e.g. , , , , , , NOTE

e.g. , , , , , , ,

- Rules for judging sub-categories of numeral: (1) *mx* can be followed by either measure word or *mw*; (2) *mg* can be followed by measure word, but cannot be followed by *mw*; (3) *mm* can be followed by measure word.
- 10. measure word (11 tags)
- 10.1. nominal measure word
- 10.1.1. qng: universal measure word ""
- e.g. \_\_\_\_\_, 10.1.2. <u>qnm</u>: unit measure word
- e.g. , , , , , , , 10.1.3. <u>qnc</u>: uncertain measure word e.g. , ,
- 10.1.4. <u>qnk</u>: type measure word e.g. ,
- 10.1.5. <u>qns</u>: individual measure word e.g. , , , ,
- 10.1.6. <u>qnu</u>: group measure word e.g. , , , , ,
- 10.1.7. <u>qnt</u>: temporary nominal measure word
- e.g. \_ , \_ , \_ , \_ 10.2. predicative measure word
- 10.2.1. <u>qv</u>: general predicative measure word e.g. \_, \_, \_, \_, \_,
- 10.2.2. <u>qvt</u>: temporary predicative measure word e.g. \_, \_, \_, \_\_,
- 10.3. <u>qt</u>: temporal measure word e.g. , , , , , , , , ,
- 10.4. <u>qc</u>: compound measure word e.g. , ,

11. pronoun (6 tags) 11.1. rn: nominal pronoun e.g. , , , , , , , , 11.2. rt: temporal pronoun e.g. 11.3. rs: locative pronoun e.g. , , , , 11.4. rv: predicative pronoun , , , e.g. 11.5. rb: distinctive pronoun e.g. , , , 11.6. rd: adverbial pronoun e.g. , , , , , , 12. preposition (4 tags) 12.1. pba: prepositions " 12.2. <u>pbe</u>: prepositions " 12.3. pza: prepositions " / " 12.4. pg: general preposition e.g. , , , , , 13. adverbial (4 tags) 13.1. db: pre-negative adverbial e.g. , , , , 13.2. dn: negative adverbial e.g. , , , , , 13.3. dd: degree adverbial e.g. , , , , , , , 13.4. dr: adverbial, others e.g. , , , , , , , , , , , 14. <u>c</u>: conjunction (1 tag) e.g. , , , , , , , , , 15. auxiliary (12 tags) 15.1. structural auxiliary 15.1.1. usd: structural auxiliary " " 15.1.2. <u>usz</u>: structural auxiliary "" 15.1.3. <u>uss</u>: structural auxiliary "" 15.1.4. <u>usi</u>: structural auxiliary "" 15.1.5. usf: structural auxiliary " " 15.2. temporal auxiliary 15.2.1. <u>utl</u>: temporal auxiliary "" 15.2.2. <u>utz</u>: temporal auxiliary "" 15.2.3. <u>utg</u>: temporal auxiliary "" 15.3. up: potential complement auxiliary e.g. \_ , \_ 15.4. ub: comparative auxiliary e.g. ( ) 15.5. <u>uo</u>: omission auxiliary " " " " 15.6. ur: auxiliary, others \_\_\_\_, / .....\_\_, ......\_\_ e.g. NOTE " should be tagged as usf. 16. y: modal word (1 tag) e.g. , , ,

17. <u>e</u> interjection (1 tag) e.g. , , , ,

18. o onomatopoeia (1 tag) e.g. , , , 19. affix (4 tags) 19.1. kh: nominal prefix e.g. , 19.2. kn: nominal suffix e.g. , , 19.3. kv: predicative suffix e.g. 19.4. ki: inflectional suffix e.g. \_, 20. idiom (3 tags) 20.1. in: nominal idiom e.g. 20.2. iv: predicative idiom e.g. , , 20.3. id: adverbial idiom e.g. • 21. <u>l</u>: linking phrase (1 tag) e.g. , 22. miscellaneous (24 tags) 22.1. zi: any Chinese character which cannot be classified into categories above e.g. "\_\_\_\_\_ 22.2. xm: digital string e.g. 123, 123000, 1.786

22.3. <u>xe</u>: character string(mixture of English character and digital number)

e.g. IBM, CAI, Prolog, Computer, IBM-PC, WPS97 22.4. <u>xg</u>: any string, others

e.g. y=x+1, v=s/t, ,

22.5. punctuation(total 20 tags): every punctuation defined in the national standard on punctuation (P.R. China) is regarded as a distinct sub-category. And, each tag of a punctuation is simply represented by itself.

Here, we show some pieces of texts out of each genre of *HuaYu* for reader's reference.

#### Literature

|rn||pg||rn||vgs||" rn |dn vgs |f |vgn |rd |a |usd |a |ng | dd vi vgp ng dr vgi rs dr vgn ng rn vgs rn ng dr pg rn nps usd a c a utl | |rn |f ng usd a |ng |dd pg |ng |f |vgi |vgc |dr |vi |pg |mx qng ng usd vgn a usd ng |ng | |pza |a |mx |qng |ng |f |vgi |vgc |usd | c vi ng c ng pg |ng |dr |a | "|"

### News

xm qt npc ng ng pza nps |xm |mab |qns |ng |pg |rn ng vgi usd |ng |f |vgn |ng | |mx |qns |dr |vgn |npf | pg |rn |vgi |usf |dd |a |usd |ng | dr usi vgn |ng |s |usd |ng | "|" |mx |qns ng "|" |pg |rn |vgi |usf |a |usd |ng |vgi ng vgs

| |c |pg |rn |vgi |usf |a |usd |ng |va |vgv |vgi | "|"

#### Academy

ng ng ng | ng vi ng - 1 ng vgx c vgx uss vgi C rn pza ng f usd ng dd a c dd a dr va vgv vgi |c vgi | ng vgi usd c |ng |vi |va |pg vgn a usd ng mx qng ng usd a ng ng pg vgp vgi vgc ng ng ng ng ng uo ng |ng |

#### Practical writing

dr vgn qnm xm ng npr ng ng xm xm |ng xm qnm vgn ng anm ng xm qnm xm ng xm ng vgn qnm qnm npr ng | xm ng nps |f vgn xm qnm ng az vgi

vgn ng ng ng xm qt ng ng ng ng ng ng dr vgi ng npx |ng @|@ ng xm xe xm xg |xm

# 4. Concluding Remarks

Though we have been doing very carefully in the procedure of conducting HuaYu, there still remain some unsolved or not satisfactorily solved issues, for example, annotation consistency. The lessons learned from previous stages of establishing HuaYu will be surely beneficial to the improvement of it. Several ongoing research projects on Chinese computing at Tsinghua University, including one with destination of implementing a high-performance Chinese word segmentation and part-of-speech tagging system for unrestricted texts and another one with destination of setting up a Chinese tree bank composed of 10,000 sentences (as mentioned in section 2.2), are exploiting HuaYu intensively to facilitate the work as retrieving linguistic evidence systematically and then making further analysis (statistical analysis in most cases) based on these rich observations. In the future, we'll update HuaYu continually, making the annotated corpus more consistent and more satisfactory. Readers who are interested in HuaYu may find additional information on our web page (unfortunately it is only in GB code now, so browser with Chinese GB is needed): а http://dns.itsl.tsinghua.edu.cn/ainlp/source.htm. The research in natural language processing and Chinese computing at Tsinghua University is also accessible on http://dns.itsl.tsinghua.edu.cn/ainlp/index.htm. page The English version of our web site is under construction currently.

## 5. References

- Chen, K.J., Huang, C.R., Chang, L.P., Hsu, H.L., 1996. SINACA CORPUS: Design methodology for balanced corpora. *PACLIC11 Proceedings*, 167-176.
- Hu, J.F., Yu, S.W., 2000. The multi-layer language knowledge base of Chinese NLP. *LREC2000 Proceedings*.
- Sun, M.S., Shen, D.Y., Huang, C.N., 1997. CSeg&Tag1.0: A practical word segmenter and POS tagger for Chinese texts, ANLP97 Proceedings, 119-126.
- Sun, M.S., Shen, D.Y., Benjamin, K.T., 1998. Chinese word segmentation without using lexicon and handcrafted training data. COLING-ACL98 Proceedings, 1265-1271.
- Zhou, Q., Sun, M.S., 2000. Build a Chinese tree bank as the test suite for Chinese parser. Proceedings of the workshop MAL'99 (Multi-lingual Information Processing and Asian Language Processing), 32-36.