IMPLEMENTING A SUBCATEGORIZED PROBABILISTIC DEFINITE CLAUSE GRAMMAR FOR VIETNAMESE SENTENCE PARSING

Dang Tuan Nguyen, Kiet Van Nguyen, Tin Trung Pham

Faculty of Computer Science, University of Information Technology,
Vietnam National University – Ho Chi Minh City,
Ho Chi Minh City, Vietnam

{ntdang, nvkiet, pttin}@nlke-group.net

ABSTRACT

In this paper, we introduce experiment results of a Vietnamese sentence parser which is built by using the Chomsky's subcategorization theory and PDCG (Probabilistic Definite Clause Grammar). The efficiency of this subcategorized PDCG parser has been proved by experiments, in which, we have built by hand a Treebank with 1000 syntactic structures of Vietnamese training sentences, and used different testing datasets to evaluate the results. As a result, the precisions, recalls and F-measures of these experiments are over 98%.

KEYWORDS

Probabilistic Context-Free Grammar, Probabilistic Definite Clause Grammar, Parsing, Subcategorization

1. INTRODUCTION

The PCFG (Probabilistic Context-Free Grammar) [1], [2], [3], and [4] has been applied for developing some Vietnamese parsers as [5], [6], and [7]. All of these mentioned Vietnamese parsers use un-subcategorized PCFG.

In this research, we are interested in applying the Chomsky's subcategorization theory [8] and PDCG (Probabilistic Definite Clause Grammar) [9], [10], [11], and [12] to implement a subcategorized PDCG parser which allows analyzing effectively simple Vietnamese sentences. To implement this parser, we define our Vietnamese subcategorized PDCG grammar based on our set of sub-categorical and phrasal tags, and syntactic rules defined on these tags.

We also develop a Treebank for training this subcategorized PDCG parser. The Vietnamese sentences in our Treebank are syntactically analyzed and tagged by hand.

DOI : 10.5121/ijnlc.2013.2401

2. SUB-CATEGORICAL AND PHRASAL TAGS

2.1. Definition of Subcategorical Tags

- **Nominal Tags**

Nouns are divided into 9 sub-categories as presented in Table 1.

Table 1. Sub-categories of nouns

No.	Nominal groups	Non-terminals	Adjectival tags
1	Common nouns	n	N
2	Abbreviation nouns	n_abbr	N_ABBR
3	Currency nouns	n_currency	N_CURRENCY
4	English nouns	n_eng	N_ENG
5	Idiomatic nouns	n_idiom	N_IDIOM
6	Proper nouns	n_prop	N_PROP
7	Temporal nouns	n_time	N_TIME
8	Title nouns	n_title	N_TITLE
9	Unit nouns	n_unit	N_UNIT

- **Verbal tags**

Verbs are divided into 8 sub-categories as presented in Table 2.

Table 2. Sub-categories of verbs

No.	Verbal groups	Non-terminals	Adjectival tags
1	Ordinary verbs	v	V
2	Passive verbs	v_bi	V_BI
3	Motion verbs	v_di	V_DI
4	Acquisition verbs	v_duoc	V_DUOC
5	English verbs	v_eng	V_ENG
6	Idiomatic verbs	v_idiom	V_IDIOM
7	Là (to be)	v_la	V_LA
8	Modal verbs	v_modal	V_MODAL

- **Adjectival tags**

Adjectives are divided into 10 sub-categories as presented in Table 3.

Table 3. Sub-categories of adjectives

No.	Adjectival groups	Non-terminals	Adjectival tags
1	Qualitative adjective	adj	ADJ
2	English adjective	adj_eng	ADJ_NUM
3	Idiomatic adjective	adj_idiom	ADJ_PERCENT
4	Measurement adjective	adj_measure	ADJ_QUANT
5	Numeric adjective	adj_num	ADJ_NUM
6	Ordinal adjective	adj_order	ADJ_ORDER

7	Percentage adjective	adj_percent	ADJ_PERCENT
8	Quantitative adjective	adj_quant	ADJ_QUANT
9	Year adjective	adj_year	ADJ_YEAR
10	Definite adjective	adj_dem	ADJ_DEM

- **Adverbial tags**

Adverbs are divided into 7 sub-categories as presented in Table 4.

Table 4. Sub-categories of adverbs

No.	Adverbial groups	Non-terminals	Adverbial tags
1	Ordinary adverb	adv	ADV
2	Estimative adverb	adv_est	ADV_EST
3	Adverb of frequency	adv_freq	ADV_FREG
4	Negative adverb	adv_neg	ADV_NEG
5	Ordinary adverb	adv_order	ADV_ORDER
6	Adverb of time	adv_tense	ADV_TENSE
7	Special adverb	adv_sp	ADV_SP

- **Prepositional tags**

Prepositions are divided into 3 groups, and ungrouped 17 prepositions. See in Table 5.

Table 5. Groups of prepositions

No.	Prepositional groups	Non-terminals	Prepositional tags
1	Preposition of cause	prep_cause	PREP_CAUSE
2	Preposition of direction	prep_direct	PREP_DIRECT
3	Preposition of location	prep_location	PREP_location
4	Bằng	prep_bang	PREP_BANG
5	Cho	prep_cho	PREP_CHO
6	Của	prep_cua	PREP_CUA
7	Cùng	prep_cung	PREP_CUNG
8	Để	prep_de	PREP_DE
9	Khi	prep_khi	PREP_KHI
10	Khỏi	prep_khoi	PREP_KHOI
11	Không	prep_khong	PREP_KHONG
12	Nếu	prep_neu	PREP_NEU
13	Qua	prep_qua	PREP_QUA
14	Sau	prep_sau	PREP_SAU
15	Trong	prep_trong	PREP_TRONG
16	Trước	prep_truoc	PREP_TRUOC
17	Từ	prep_tu	PREP_TU
18	Vào	prep_vao	PREP_VAO
19	Về	prep_ve	PREP_VE
20	Với	prep_voi	PREP_VOI

- **Conjunctional tags**

Table 6. Groups of prepositions

No.	Conjunctional groups	Non-terminals	Conjunctional tags
1	và, "-", ","	conj	CONJ

- **Special tags**

In Vietnamese, there are words that always precede a noun or an adjective to modify for the noun or adjective, e.g. "cố", "cựu", "phó", "siêu", "tân", … We arrange these special words in a group.

Table 7. Groups of special words

No.	Special groups	Non-terminals	Special tags
1	Special words	sp_word	SP_WORD

2.2. Phrasal tags

- **Verbal phrase tags**

Verbal phrase are divides in to 5 groups as shown in Table 8.

Table 8. Groups of verbal phrases

No.	Verbal phrase groups	Non-terminals	Verbal phrase tags
1	Verbal phrase having a intransitive verb	vp1	VP1
2	Verbal phrase having a transitive verb and its direct object	vp2	VP2
3	Verbal phrase having a transitive verb and its indirect object	vp3	VP3
4	Verbal phrase having a transitive verb and its direct and indirect object	vp4	VP4
5	General verbal phrase	vp	VP

- **Nominal phrase tags**

Noun phrases are divided into 8 groups, as presented in Table 9.

Table 9. Groups of nominal phrases

No.	Nominal phrase groups	Non-terminals	Nominal phrase tags
1	General noun phrase	np	NP
2	Noun phrases have two components linked together by connected words or hyphen	np_conj	NP_CONJ
3	Currency noun phrases	np_currency	NP_CURRENCY
4	Noun phrases contains one, or two, or three, or four, or five noun(s)	np_n np_nn np_nnn np_nnnn np_nnnnn	NP_N NP_NN NP_NNN NP_NNNN NP_NNNNN

5	Noun phrases contain one, or two, or three noun(s) that precede a preposition	np_npp np_nnpp np_nnnpp	NP_NPP NP_NNPP NP_NNNPP
6	Noun phrases of pronoun	np_pn	NP_PN
7	Noun phrases of proper name	np_prop	NP_PROP
8	Noun phrases of time	np_time	NP_TIME

- **Prepositional phrase tags**

Prepositional phrases are divided into 15 groups based on preposition, as presented in Table 10.

Table 10. Groups of prepositional phrases

No.	Prepositional groups	Non-terminals	Prepositional tags
1	Prepositional phrases contain the preposition “bằng”	pp_bang	PP_BANG
2	Prepositional phrases of cause	pp_cause	PP_CAUSE
3	Prepositional phrases contain the preposition “cho”	pp_cho	PP_CHO
4	Prepositional phrases contain the preposition “của”	pp_cua	PP_CUA
5	Prepositional phrases contain the preposition “cùng”	pp_cung	PP_CUNG
6	Prepositional phrases contain a preposition of direction	pp_direct	PP_DIRECT
7	Prepositional phrases contain the preposition “khi”	pp_khi	PP_KHI
8	Prepositional phrases contain the preposition “không”	pp_khong	PP_KHONG
9	Prepositional phrases of location	pp_location	PP_LOCATION
10	Prepositional phrases contain the preposition “qua”	pp_qua	PP_QUA
11	Prepositional phrases contain the preposition “sau”	pp_sau	PP_SAU
12	Prepositional phrases contain the preposition “trong”	pp_trong	PP_TRONG
13	Prepositional phrases contain the preposition “trước”	pp_truoc	PP_TRUOC
14	Prepositional phrases contain the preposition “vào”	pp_vao	PP_VAO
15	Prepositional phrases contain the preposition “về”	pp_ve	PP_VE

- **Adjectival phrase tags**

Adjectival phrases are divided into 3 groups and presented in Table 11.

Table 11. Groups of adjectival phrases

No.	Adjectival phrase groups	Non-terminals	Adjectival phrase tags
1	Adjectival phrase of quality	adjp	ADJP
2	Adjectival phrase of number	adjp_num	ADJP_NUM
3	Adjectival phrase of measurement	adjp_measure	ADJP_MEASURE

3. SYNTACTIC RULES OF PHRASES

The probabilities of phrasal structure rules are calculated with 1000 Vietnamese training sentences in our TreeBank described in the experiments.

3.1. Nominal phrase

- **Nominal phrase of one noun**

Nominal phrase NP_N is formed by a noun or/and an adjective modifying for noun.

Table 12. Syntactic rules of NP_N phrases

No.	Rules	Probabilities
1	NP_N → ADJ_PRE N	0.005714
2	NP_N → N	0.809524
3	NP_N → N ADJP	0.179048
4	NP_N → N ADJP_MEASURE	0.001905
5	NP_N → N ADJP_NUM	0.003810

- **Nominal phrase of two nouns**

Nominal phrase NP_NN is formed by two nouns and without adjuncts.

Table 13. Syntactic rules of NP_NN phrases

No.	Rules	Probabilities
1	NP_NN → N N	1.0

- **Nominal phrase of three nouns**

Nominal phrase NP_NNN is formed by three nouns and without complements.

Table 14. Syntactic rules of NP_NNN phrases

No.	Rules	Probabilities
1	NP_NNN → N N N	1.0

- **Noun phrase of four nouns**

Nominal phrase NP_NNNN formed by four nouns and without complements.

Table 15. Syntactic rules of NP_NNNN phrases

No.	Rules	Probabilities
1	NP_NNNN → N N N N	1.0

- **Nominal phrase of five nouns**

Nominal phrase NP_NNNNN formed by five nouns and without complements.

Table 16. Syntactic rules of NP_NNNNN phrases

No.	Rules	Probabilities
1	NP_NNNNN → N N N N N	1.0

- **Nominal phrase of proper nouns**

Table 17. Syntactic rules of proper nouns

No.	Rules	Probabilities
1	NP_PROP → N_PROP	0.994715
2	NP_PROP → N_PROP N_PROP	0.002114
3	NP_PROP → N_TITLE N_PROP	0.003171

- **Nominal phrase of time**

Table 18. Syntactic rules for nominal phrases of time

No.	Rules	Probabilities
1	NP_TIME → ADJ_PRE N_TIME	0.045455
2	NP_TIME → ADJ_PRE N_TIME ADJ_DEM	0.022727
3	NP_TIME → ADJ_QUANT N_TIME	0.022727
4	NP_TIME → ADJ_QUANT N_TIME ADJP_NUM	0.022727
5	NP_TIME → ADJP_NUM N_TIME	0.159091
6	NP_TIME → ADJP_NUM N_TIME ADJ_DEM	0.022727
7	NP_TIME → ADJP_NUM N_TIME N	0.022727
8	NP_TIME → N N_TIME	0.022727
9	NP_TIME → N_TIME	0.090909
10	NP_TIME → N_TIME ADJ	0.045455
11	NP_TIME → N_TIME ADJ_DEM	0.045455
12	NP_TIME → N_TIME ADJ_ORDER	0.022727
13	NP_TIME → N_TIME ADJP_NUM	0.045455
14	NP_TIME → N_TIME DATE	0.022727
15	NP_TIME → N_TIME N	0.318182
16	NP_TIME → N_TIME N N	0.068182

- **Nominal phrase of currency**

Table 19. Syntactic rules for nominal phrases of currency

No.	Rules	Probabilities
1	NP_CURRENCY → ADJP_NUM N_CURRENCY	1.0

- **Nominal phrase containing conjunctions**

Table 20. Syntactic rules of nominal phrases containing conjunctions

No.	Rules	Probabilities
1	NP_CONJ → NP_N CONJ NP_N	0.066667
2	NP_CONJ → NP_NN CONJ NP_NN	0.038095

3	NP_CONJ → NP_PROP CONJ NP_N	0.019048
4	NP_CONJ → NP_PROP CONJ NP_PROP	0.819048
5	NP_CONJ → NP_PROP CONJ NP_PROP CONJ NP_PROP	0.057143

- **Nominal phrase np_xpp is formed by np_n and pp_**

Table 21. Syntactic rules of np_xpp

No.	Rules	Probabilities
1	NP_NPP → NP_N PP_CUA	0.688525
2	NP_NPP → NP_N PP_KHONG	0.114754
3	NP_NPP → NP_N PP_LOCATION	0.098361
4	NP_NPP → NP_N PP_TRONG	0.032787
5	NP_NPP → NP_N PP_VE	0.065574

- **Nominal phrase np_nnpp is formed by np_nn and pp_**

Table 22. Syntactic rules of np_nnpp

No.	Rules	Probabilities
1	NP_NNPP → NP_NN PP_CHO	0.024390
2	NP_NNPP → NP_NN PP_CUA	0.512195
3	NP_NNPP → NP_NN PP_LOCATION	0.390244
4	NP_NNPP → NP_NN PP_QUA	0.024390
5	NP_NNPP → NP_NN PP_TRONG	0.048780

- **Nominal phrase np_nnnpp is formed by np_nnn and pp_**

Table 23. Syntactic rules of np_nnnpp

No.	Rules	Probabilities
1	NP_NNNPP → NP_NNN PP_CUA	0.6
2	NP_NNNPP → NP_NNN PP_LOCATION	0.3
3	NP_NNNPP → NP_NNN PP_TRONG	0.1

- **Nominal phrase np_nnnnpp is formed by np_nnnn and pp_**

Table 24. Syntactic rules of np_nnnnpp

No.	Rules	Probabilities
1	NP_NNNNPP → NP_NNNN PP_CUA.	1.0

- **Nominal phrase**

Table 25. Syntactic rules of general nominal phrases

No.	Rules	Probabilities
1	NP → ADJ_PERCENT NP_N	4.46E-04
2	NP → ADJ_PERCENT NP_NN	8.93E-04
3	NP → ADJ_QUANT NP_N	0.003571
4	NP → ADJ_QUANT NP_N PP_CUA	8.93E-04
5	NP → ADJ_QUANT NP_N PP_LOCATION	4.46E-04
6	NP → ADJ_QUANT NP_N PP_TRONG	8.93E-04
7	NP → ADJ_QUANT NP_N PP_VOI	8.93E-04
8	NP → ADJ_QUANT NP_NN	0.002232
9	NP → ADJ_QUANT NP_NN PP_CUA	8.93E-04
10	NP → ADJ_QUANT NP_NN PP_TRONG	0.001339
11	NP → ADJ_QUANT NP_NNN	0.001339
12	NP → ADJ_QUANT NP_NNNN	4.46E-04
13	NP → ADJP_MEASURE NP_N	4.46E-04
14	NP → ADJP_MEASURE NP_NN	4.46E-04
15	NP → ADJP_NUM NP_N	0.012054
16	NP → ADJP_NUM NP_N NP_TIME	4.46E-04
17	NP → ADJP_NUM NP_N PP_CUA	8.93E-04
18	NP → ADJP_NUM NP_N PP_TRONG	4.46E-04
19	NP → ADJP_NUM NP_NN	0.0125
20	NP → ADJP_NUM NP_NN ADJP	4.46E-04
21	NP → ADJP_NUM NP_NN PP_CUA	8.93E-04
22	NP → ADJP_NUM NP_NN PP_LOCATION	4.46E-04
23	NP → ADJP_NUM NP_NNN	0.002679
24	NP → ADJP_NUM NP_NNNN	0.001339
25	NP → ADJP_NUM SP_WORD NP_N	4.46E-04
26	NP → ADJP_NUM SP_WORD NP_NNN	4.46E-04
27	NP → N ADJ_QUANT NP_N	4.46E-04
28	NP → N ADJP_NUM NP_N	0.002679
29	NP → N ADJP_NUM NP_NN	4.46E-04
30	NP → N NP_CONJ	0.008482
31	NP → N NP_CURRENCY	0.001786
32	NP → N NP_CURRENCY ADJP_MEASURE	4.46E-04
33	NP → N NP_N	4.46E-04
34	NP → N NP_PROP	4.46E-04
35	NP → N SP_WORD NP_N	4.46E-04
36	NP → N SP_WORD NP_N ADJP	4.46E-04
37	NP → N SP_WORD NP_NN	8.93E-04

38	NP → NP_CONJ	0.036161
39	NP → NP_CONJ ADJP PP_LOCATION	4.46E-04
40	NP → NP_CONJ PP_TRONG	4.46E-04
41	NP → NP_CURRENCY	0.003125
42	NP → NP_N	0.165625
43	NP → NP_N NP_PROP	0.001339
44	NP → NP_NN	0.215179
45	NP → NP_NN ADJP	0.014732
46	NP → NP_NN ADJP PP_CUA	0.002232
47	NP → NP_NN ADJP_NUM	4.46E-04
48	NP → NP_NN ADJP_NUM NP_N	4.46E-04
49	NP → NP_NN NP_CONJ	0.001339
50	NP → NP_NN NP_CURRENCY PP_CUA	4.46E-04
51	NP → NP_NN NP_TIME	4.46E-04
52	NP → NP_NN NP_TIME PP_LOCATION	4.46E-04
53	NP → NP_NN SP_WORD NP_NN	4.46E-04
54	NP → NP_NNN	0.061161
55	NP → NP_NNN ADJP	0.001339
56	NP → NP_NNN ADJP PP_CUA	4.46E-04
57	NP → NP_NNN NP_TIME	4.46E-04
58	NP → NP_NNNN	0.009821
59	NP → NP_NNNNN	0.001339
60	NP → NP_NNNNPP	8.93E-04
61	NP → NP_NNNPP	0.004464
62	NP → NP_NNPP	0.017857
63	NP → NP_NNPP NP_TIME	4.46E-04
64	NP → NP_NPP	0.026339
65	NP → NP_NPP PP_LOCATION	8.93E-04
66	NP → NP_PN	4.46E-04
67	NP → NP_PROP	0.331696
68	NP → NP_PROP ADJP	4.46E-04
69	NP → NP_PROP ADJP PP_LOCATION	8.93E-04
70	NP → NP_PROP NP_TIME	4.46E-04
71	NP → NP_PROP PP_CUA PP_LOCATION	4.46E-04
72	NP → NP_PROP PP_TRUOC	4.46E-04
73	NP → NP_TIME	0.011161
74	NP → NP_TIME ADJP PP_LOCATION	4.46E-04
75	NP → NP_TIME PP_CUA PP_LOCATION	4.46E-04
76	NP → NP_TIME PP_LOCATION	4.46E-04
77	NP → SP_WORD NP_N	0.001786
78	NP → SP_WORD NP_N PP_CUA	4.46E-04

79	NP → SP_WORD NP_NN	0.013839
80	NP → SP_WORD NP_NN PP_LOCATION	4.46E-04
81	NP → SP_WORD NP_NNN	0.001339
82	NP → SP_WORD NP_NNNN	8.93E-04

3.2. Verbal phrase

- **VP1: Verbal phrases only contain intransitive verb**

Table 26. Syntactic rules of VP1

No.	Rules	Probabilities
1	VP1 → ADV V	0.040404
2	VP1 → ADV_NEG V	0.020202
3	VP1 → ADV_TENSE V	0.050505
4	VP1 → ADV_TENSE V ADV	0.020202
5	VP1 → V	0.737374
6	VP1 → V ADV	0.121212
7	VP1 → V ADV_SP	0.010101

- **VP2: Verbal phrases with a direct object of transitive verb**

Table 27. Syntactic rules of VP2

No.	Rules	Probabilities
1	VP2 → ADV V NP	0.044444
2	VP2 → ADV V_LA NP	0.001852
3	VP2 → ADV_FREQ V_LA NP	0.001852
4	VP2 → ADV_NEG V NP	0.012963
5	VP2 → ADV_NEG V_MODAL NP	0.001852
6	VP2 → ADV_TENSE ADV V NP	0.001852
7	VP2 → ADV_TENSE ADV_NEG V NP	0.001852
8	VP2 → ADV_TENSE V ADV NP	0.005556
9	VP2 → ADV_TENSE V NP	0.068519
10	VP2 → V ADV NP	0.044444
11	VP2 → V NP	0.790741
12	VP2 → V NP ADV	0.011111
13	VP2 → V_DUOC NP	0.001852
14	VP2 → V_LA NP	0.011111

- **VP3: Verbal phrases with an indirect object of transitive verb**

Table 28. Syntactic rules of VP3

No.	Rules	Probabilities
1	VP3 → ADV V PREP_CAUSE NP	0.009479
2	VP3 → ADV V PREP_TRUOC NP	0.009479
3	VP3 → ADV V PREP_VE NP	0.009479
4	VP3 → ADV V ADV PREP_CAUSE VP1	0.004739
5	VP3 → ADV_NEG ADV V PREP_LOCATION NP	0.004739
6	VP3 → ADV_NEG V PREP_CHO NP	0.004739
7	VP3 → ADV_NEG V PREP_TRONG NP	0.004739
8	VP3 → ADV_NEG V PREP_VOI NP	0.009479
9	VP3 → ADV_TENSE V PREP_LOCATION NP	0.023697
10	VP3 → ADV_TENSE V PREP_TRONG NP	0.014218
11	VP3 → ADV_TENSE V PREP_VE NP	0.004739
12	VP3 → ADV_TENSE V PREP_VOI NP	0.004739
13	VP3 → V ADV PREP_CAUSE NP	0.014218
14	VP3 → V ADV PREP_CAUSE V_BI VP1	0.004739
15	VP3 → V ADV PREP_DE VP2	0.004739
16	VP3 → V ADV PREP_KHI V_BI VP3	0.004739
17	VP3 → V ADV PREP_KHI VP3	0.004739
18	VP3 → V ADV PREP_LOCATION NP	0.014218
19	VP3 → V ADV PREP_QUA NP	0.004739
20	VP3 → V ADV PREP_SAU NP	0.004739
21	VP3 → V ADV PREP_TRONG NP	0.004739
22	VP3 → V ADV PREP_TRUOC NP	0.004739
23	VP3 → V ADV PREP_VE NP	0.004739
24	VP3 → V ADV_SP PREP_VE NP	0.004739
25	VP3 → V PREP_BANG NP	0.014218
26	VP3 → V PREP_CAUSE NP	0.113744
27	VP3 → V PREP_CAUSE V_BI VP1	0.004739
28	VP3 → V PREP_CAUSE V_BI VP2	0.004739
29	VP3 → V PREP_CAUSE VP1	0.018957
30	VP3 → V PREP_CAUSE VP2	0.033175
31	VP3 → V PREP_CHO NP	0.023697
32	VP3 → V PREP_CUNG NP	0.028436
33	VP3 → V PREP_DE VP1	0.014218
34	VP3 → V PREP_DE VP2	0.009479
35	VP3 → V PREP_DIRECT NP	0.094787
36	VP3 → V PREP_KHI VP1	0.009479
37	VP3 → V PREP_KHOI NP	0.004739

38	VP3 → V PREP_LOCATION NP	0.156398
39	VP3 → V PREP_QUA NP	0.009479
40	VP3 → V PREP_SAU NP	0.014218
41	VP3 → V PREP_TRONG NP	0.047393
42	VP3 → V PREP_TRUOC NP	0.018957
43	VP3 → V PREP_TU NP	0.004739
44	VP3 → V PREP_VE NP	0.151659
45	VP3 → V PREP_VE VP2	0.004739
46	VP3 → V PREP_VOI NP	0.009479
47	VP3 → V CONJ V ADV PREP_LOCATION NP	0.004739
48	VP3 → V CONJ V PREP_VOI NP	0.004739
49	VP3 → V_DI ADV PREP_TU NP	0.004739
50	VP3 → V_DI PREP_DIRECT NP	0.004739
51	VP3 → V_LA PREP_CAUSE NP	0.009479

- **VP4: Verbal phrases with a direct object and an indirect object of transitive verb**

Table 29. Syntactic rules of VP4

No.	Rules	Probabilities
1	VP4 → ADV V NP PREP_DIRECT NP	0.013245
2	VP4 → ADV_NEG V NP PREP_KHOI NP	0.006623
3	VP4 → ADV_TENSE V ADV NP PREP_TU NP	0.006623
4	VP4 → ADV_TENSE V NP PREP_CAUSE NP	0.006623
5	VP4 → ADV_TENSE V NP PREP_CHO NP	0.013245
6	VP4 → ADV_TENSE V NP PREP_CUNG NP	0.006623
7	VP4 → ADV_TENSE V NP PREP_DIRECT NP	0.052980
8	VP4 → V ADV NP PREP_DE VP3	0.006623
9	VP4 → V ADV NP PREP_CHO NP	0.006623
10	VP4 → V ADV NP PREP_DIRECT NP	0.013245
11	VP4 → V NP PREP_CAUSE ADV_NEG V_DUOC VP1	0.006623
12	VP4 → V NP PREP_CAUSE VP1	0.006623
13	VP4 → V NP PREP_CAUSE VP2	0.019868
14	VP4 → V NP PREP_DE VP1	0.006623
15	VP4 → V NP PREP_DE VP2	0.052980
16	VP4 → V NP PREP_DIRECT VP1	0.006623
17	VP4 → V NP PREP_KHI VP2	0.019868
18	VP4 → V NP PREP_BANG NP	0.013245
19	VP4 → V NP PREP_CAUSE NP	0.072848
20	VP4 → V NP PREP_CHO NP	0.225166
21	VP4 → V NP PREP_CUNG NP	0.006623
22	VP4 → V NP PREP_DIRECT ADV NP	0.006623
23	VP4 → V NP PREP_DIRECT NP	0.172185

24	VP4 → V NP PREP_KHOI NP	0.019868
25	VP4 → V NP PREP_QUA NP	0.006623
26	VP4 → V NP PREP_TRONG NP	0.039735
27	VP4 → V NP PREP_TU NP	0.019868
28	VP4 → V NP PREP_VE NP	0.139073
29	VP4 → V NP PREP_CAUSE V VP2	0.006623
30	VP4 → V_BI NP PREP_CAUSE NP	0.006623
31	VP4 → V_DUOC NP PREP_CHO VP1	0.006623
32	VP4 → V_LA NP PREP_CHO NP	0.006623

- **General verbal phrases**

Table 30. Syntactic rules of general verbal phrase

No.	Rules	Probabilities
1	VP → ADV V VP1	0.008734
2	VP → ADV V VP2	0.021834
3	VP → ADV V_BI VP2	0.008734
4	VP → ADV V_BI VP3	0.004367
5	VP → ADV V_DI VP1	0.008734
6	VP → ADV V_DI VP2	0.004367
7	VP → ADV V_MODAL VP2	0.004367
8	VP → ADV_NEG V_DUOC VP2	0.008734
9	VP → ADV_NEG V_MODAL V VP1	0.004367
10	VP → ADV_NEG V_MODAL V_MODAL VP2	0.004367
11	VP → ADV_TENSE V VP2	0.008734
12	VP → ADV_TENSE V_BI VP1	0.004367
13	VP → ADV_TENSE V_DUOC VP1	0.008734
14	VP → ADV_TENSE V_DUOC VP2	0.004367
15	VP → V ADV VP2	0.004367
16	VP → V ADV_SP VP1	0.004367
17	VP → V ADV_SP VP2	0.004367
18	VP → V ADV_SP VP4	0.004367
19	VP → V V_BI VP2	0.004367
20	VP → V VP1	0.048035
21	VP → V VP2	0.270742
22	VP → V VP3	0.026201
23	VP → V VP4	0.109170
24	VP → V V VP2	0.008734
25	VP → V_BI V VP1	0.004367
26	VP → V_BI V VP2	0.004367
27	VP → V_BI V VP3	0.008734
28	VP → V_BI V VP4	0.004367

29	VP → V_BI VP1	0.052402
30	VP → V_BI VP2	0.034934
31	VP → V_BI VP3	0.065502
32	VP → V_BI VP4	0.004367
33	VP → V_DI VP2	0.026201
34	VP → V_DI VP4	0.004367
35	VP → V_DUOC V VP1	0.004367
36	VP → V_DUOC V VP2	0.017467
37	VP → V_DUOC VP1	0.026201
38	VP → V_DUOC VP2	0.004367
39	VP → V_DUOC VP3	0.008734
40	VP → V_MODAL V VP2	0.004367
41	VP → V_MODAL V_BI VP1	0.004367
42	VP → V_MODAL V_DUOC VP1	0.004367
43	VP → V_MODAL VP1	0.013100
44	VP → V_MODAL VP2	0.096070
45	VP → V_MODAL VP3	0.004367
46	VP → V_MODAL VP4	0.013100

3.3. Prepositional phrase PP_

Table 31. Syntactic rules of PP_

No.	Rules	Probabilities
1	PP_CHO → PREP_CHO NP	1.0
2	PP_CUA → PREP_CUA NP	1.0
3	PP_DIRECT → PREP_DIRECT NP	1.0
4	PP_KHONG → PREP_KHONG NP	1.0
5	PP_LOCATION → PREP_LOCATION NP	1.0
6	PP_QUA → PREP_QUA NP	1.0
7	PP_SAU → PREP_SAU NP	1.0
8	PP_TRONG → PREP_TRONG NP	1.0
9	PP_TRUOC → PREP_TRUOC NP	1.0
10	PP_VAO → PREP_VAO NP_TIME	1.0
11	PP_VE → PREP_VE NP	1.0
12	PP_VOI → PREP_VOI NP	1.0

3.4. Adjectival phrase

Adjectival phrase consists of three types of general adjectival phrase (adjp), adjectival phrase of number (adjp_num), adjectival phrase of number (adjp_measure).

- **General adjectival phrase (ADJP)**

Table 32. Syntactic rules of ADJP

No.	Rules	Probabilities
1	ADJP → ADJ	0.797203
2	ADJP → ADJ ADV_ORDER	0.055944
3	ADJP → ADJ ADV_ORDER NP_N	0.062937
4	ADJP → ADJ ADV_ORDER NP_NN	0.020979
5	ADJP → ADJ ADV_ORDER NP_PROP	0.006993
6	ADJP → ADJ ADV_ORDER NP_TIME	0.006993
7	ADJP → ADJ ADJ	0.006993
8	ADJP → ADJ CONJ ADJ CONJ ADJ	0.006993
9	ADJP → ADJ_ORDER	0.027972
10	ADJP → SP_WORD ADJ	0.006993

- **Adjectival phrase of number (ADJP_NUM)**

Table 33. Syntactic rules of ADJP_NUM

No.	Rules	Probabilities
1	ADJP_NUM → ADJ_NUM	0.759690
2	ADJP_NUM → ADJ_NUM ADJ_NUM	0.116279
3	ADJP_NUM → ADJ_QUANT ADJP_NUM	0.069767
4	ADJP_NUM → ADV_EST ADJP_NUM	0.054264

- **Adjectival phrase of measurement (ADJP_MEASURE)**

Table 34. Syntactic rules of ADJP_MEASURE

No.	Rules	Probabilities
1	ADJP_MEASURE → ADJ_QUANT ADJ_MEASURE	0.25
2	ADJP_MEASURE → ADJP_NUM ADJ_MEASURE	0.75

4. SYNTACTIC RULES OF SENTENCE

The probabilities of sentential structure rules are calculated with 1000 Vietnamese training sentences in our TreeBank described in the experiments.

4.1. Sentences formed by a syntagm

Table 35. Syntactic rules of sentences formed by a syntagm

No.	Rules	Probability
1	SENTENCE → NP	0.065
2	SENTENCE → VP2	0.006

3	SENTENCE → VP3	0.006
4	SENTENCE → VP4	0.01
5	SENTENCE → VP	0.002

4.2. Sentences formed by two syntagms

Table 36. Syntactic rules of sentences formed by two syntagms

No.	Rules	Probability
1	SENTENCE → NP VP1	0.037
2	SENTENCE → NP VP2	0.29
3	SENTENCE → NP VP3	0.156
4	SENTENCE → NP VP4	0.105
5	SENTENCE → NP VP	0.178
6	SENTENCE → VP2 VP2	0.003
7	SENTENCE → VP CONJ VP2	0.002
8	SENTENCE → VP4 VP3	0.001
9	SENTENCE → VP2 PP_LOCATION	0.007
10	SENTENCE → VP2 PP_SAU	0.001

4.3. Sentences formed by three syntagms

Table 37. Syntactic rules of sentences formed by three syntagms

No.	Rules	Probability
1	SENTENCE → NP VP2 PP_LOCATION	0.037
2	SENTENCE → NP VP2 PP_SAU	0.006
3	SENTENCE → NP VP2 PP_TRUOC	0.006
4	SENTENCE → NP VP2 PP_VOI	0.003
5	SENTENCE → NP VP3 PP_LOCATION	0.011
6	SENTENCE → NP VP3 PP_VOI	0.003
7	SENTENCE → NP VP4 PP_DIRECT	0.001
8	SENTENCE → NP VP4 PP_LOCATION	0.002
9	SENTENCE → NP VP PP_LOCATION	0.025
10	SENTENCE → NP VP PP_SAU	0.002
11	SENTENCE → NP VP PP_VAO	0.001
12	SENTENCE → NP VP PP_VOI	0.013
13	SENTENCE → NP VP2 NP_TIME	0.003
14	SENTENCE → NP VP3 NP_TIME	0.002
15	SENTENCE → NP VP NP_TIME	0.003
16	SENTENCE → NP VP2 VP2	0.008
17	SENTENCE → NP VP2 VP3	0.001
18	SENTENCE → NP VP2 VP	0.001
19	SENTENCE → NP VP3 VP2	0.001
20	SENTENCE → NP VP CONJ VP2	0.001
21	SENTENCE → VP2 VP PP_LOCATION	0.001

5. EXPERIMENTS

To train our subcategorized PDCG parser, we build a Treebank with 1000 simple Vietnamese training sentences, which are manually analyzed and tagged by using our Vietnamese subcategorized PDCG grammar and sub-categorical and phrasal tags. All of these Vietnamese sentences are titles of international news (from January 2011 to May 2013) which are collected and selected from the web site of VnExpress [13]. Our Treebank is based on Penn Treebank [14], in which all of the rules are represented by Sandiway Fong's Prolog formats [15].

Based on the built Treebank, the parser can extract 36 sentential rules, 309 phrasal rules and 3248 lexical rules. The probabilities of these rules are calculated by following Probabilistic Definite Clause Grammar [9], [10], [11], and [12].

The results of the experiments of the parser are presented in Table 38. We apply the evaluation method proposed by [16].

Table 38. Results of testing the subcategorized PDCG parser

Testing datasets	Number of sentences	Precision	Recall	F
Dataset 1	250	98.46	98.50	98.48
Dataset 2	500	98.42	98.44	98.43
Dataset 3	750	98.78	98.80	98.79
Dataset 4	1000	98.78	98.82	98.80

For all of experiments, the averages of precisions, recalls and F measures are over 98%.

6. CONCLUSIONS

The application of Chomsky's principle of subcategorization [8] and PDCG (Probabilistic Definite Clause Grammar) [9], [10], [11], and [12] allows enhancing the precision, recall and F measures of parsing on all of experimented Vietnamese sentences. However, building a subcategorized PDCG for Vietnamese language requires much time and linguistic complexity in defining tagset, and syntactic rules as well as building a Treebank.

In future works, we prepare to standardize the subcategorization, the tagset and syntactic rules for Vietnamese language. At the same time, the lexicon of parser will be also extended. A strong subcategorized PDCG parser will allow analyzing syntax with better precision.

REFERENCES

[1] Michael Collins, "Three generative lexicalized models for statistical parsing", Proceeding ACL '98 Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, pp. 16-23, 1997.

[2] Michael Collins, "Head-Driven Statisticcal Models for Natural Language Parsing", Journal Computational Linguistics, MIT Press, Volume. 29, No. 4, pp. 589-637, 2003.

[3] Charniak Eugene, "Statistical techniques for natural language parsing", AI Magazine, 1997.

[4] Chistopher D. Manning, Hinrich Schütze, Foundations of Statistical Natural Language Processing, MIT Press, 1999.

[5] Nguyen Quoc The, Le Thanh Huong, "Phân tích cú pháp tiếng Việt sử dụng văn phạm phi ngữ cảnh từ vựng hóa kết hợp xác suất", Proceedings of the FAIR conference, Nha Trang, Vietnam, Aug. 9-10, 2007.

[6] Hoang Anh Viet, Dinh Thi Phuong Thu, Huynh Quyet Thang, "Vietnamese Parsing Applying The PCFG Model", Proceedings of the Second Asia Pacific International Conference on Information Science and Technology, Vietnam, 2007.

[7] Đề tài VLSP. Available at: http://vlsp.vietlp.org:8080 .

[8] Noam Chomsky, *Aspects of the theory of syntax*, The M.I.T. Press, 1965.

[9] Qaiser Abbas, Nayyara Karamat, Sadia Niazi, "Development of Tree-bank Based Probabilistic Grammar for Urdu Language", International Journal of Electrical & Computer Sciences (IJECS), Vol. 9, No. 9, 2009.

[10] Parsing PCFG in Prolog. Available at: http://w3.msi.vxu.se/~nivre/teaching/statnlp/pdcg.html

[11] Assignment for PCFG Parsing. Available at: http://stp.lingfil.uu.se/~nivre/5LN437/statmet_ass2.html

[12] Gerald Gazdar, 1999. Available at: http://www.informatics.sussex.ac.uk/research/groups/nlp/gazdar/teach/nlp/

[13] VnExpress. Available at: http://www.vnexpress.net

[14] Mitchell P. Marcus, Mary Ann Marcinkiewicz, Beatrice Santorini, "Building a Large Annotated Corpus of English: The Penn Treebank", Journal Computational Linguistics - Special issue on using large corpora: II, MIT Press, Volume. 19, No. 2, pp. 313-330, 1993.

[15] Sandiway Fong, Treebank Viewer. Available at: http://dingo.sbs.arizona.edu/~sandiway/treebankviewer/index.html

[16] E. Black, "A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars", Proceedings DARPA Speech and Natural Language Workshop, Pacific Grove, Morgan Kaufmann, 1991.