A new model of Time Expressions Detection and Annotation in Vietnamese : The *hôm* case.

Philippe LAMBERT (Institut Jean Lamour, CNRS/Universite de Lorraine et GR2IV, Vinalor)
Sylviane R. SCHWER (Université Paris 13, Sorbonne Paris Cité, LIPN, CNRS UMR 7030).
Nicolas BOFFO (Praxiling/Université Montpellier 3 et MICA/Institut Polytechnique de Han oi).

***Abstract*— We describe our approach to build a new model of Time Expressions Recognition in Vietnamese, including a temporal reasoning tool in order to provide an automatic supervised learning system based on a socio-cultural linguistic approach. The Vietnamese morpheme hôm is taken for examplification of the impediments to be solved.**

Keywords: Natural Language Processing, automatic translation (Vietnamese-French), Vietnamese calendrical expressions, temporal annotation and reasoning, Viet4Nooj module.

I. INTRODUCTION

The importance of modeling temporal information is increasingly apparent in natural language applications, such as information extraction, question answering, translation or second language acquisition. Taking into account temporality expressed in texts appears as fundamental not only in a perspective of global processing of documents but also in the comprehension of the structure of a document. A lot of works have put the emphasis on the importance of the temporal expressions as modes of discursive organization. Among temporal expressions, relative calendrical expressions such as *Sunday, today, yesterday morning, the day after tomorrow, three days before* can be viewed as a part of Named Entity Resolution - they refer to calendrical periods – strongly depending on contextual or textual cues. In learning and translating perspective, the usual practice of encoding the calendrical periods designating by temporal expressions in terms of dates is far to be sufficient. Our Claim is that different expressions referring to the same period design different perspectives of the writers, which have to be kept into account during the comprehension and the translation process. For instance, Fr. *Pâques 2012* refers to the same date as *le 8 avril 2012* but the former refers to a festival day with all its religious characters, the latter to a simple Gregorian date. The same remark can be done in the Vietnamese context with the celebration doctors day (*Ngày Thầy thuốc VN*, February 27) or the Liberation Day of Sài Gòn (*Ngày Giải phóng Miền Nam*, April 30) . Some calendrical expression can have different meanings, depending on co-text or pragmatic situation. But none are able to distinguish the different meanings of *Fr. hier/aujourd'hui/demain*, that can either refer to the day-unit (1) or to the partition of a wider span of time (2).

(1) ***Hier** la pluie, **aujourd'hui** la pluie et **demain** encore la pluie.* (oral sentence, recently eared)

(2) *Ainsi, **hier** une république, **aujourd'hui** une monarchie, et **demain** encore une république.*
CHATEAUBRIAND, Essai sur les Révolutions, t. 2, 1797, p. 95.

Actual time expressions detection and annotation systems for different languages aim at computing the period refered by the expressions and provide normalizing expressions as annotations [see 1, 2, 3 among others]. For instance, using the ISO standard temporal markup

language TimeML in [1] *next Tuesday* is encapsulated inside the normalized form “two weeks from” and provides the Gregorian date associated as shown in (3).

(3) <TIMEX2 VAL="1999-08-03">two weeks from <TIMEX2 VAL="1999-07-20">next Tuesday</TIMEX2></TIMEX2>

Actually, the automatic translation systems are based on probabilitic and statistic functions. For instance, they have difficulty in differentiating between *hôm* and *ngày,* as shown in (4).

(4) Google translation translate the Fr. source *“**Aujourd'hui** nous sommes dans une situation économique difficile.”* into the *Viet*. Target *“**Hôm nay** chúng ta đang ở trong tình trạng kinh tế khó khăn”.*

Hôm nay cannot be associated in this context to ***Aujourd'hui,*** which is used as in ex. (2). One has to choose between *ngày nay* or *hiện tại*. In [2] is mentioned the difficulty in dealing correctly with temporal expressions inside probabilistic and statistic translation systems.

In Natural Language Processing community, the first implementation of some Vietnamese temporal expressions has been made inside the NooJ language environment [3]. This language is based on final automata transducers, or other said, on formal regular grammars. The Viet4NooJ (V4N) module refers to the linguistic resources created specifically for data treatments in Vietnamese[4, 5]. A first attempt to add a temporal component has been made for dates’ extraction [6]. We will get advantage of the very beginning of this task to create a multidiscipline group around temporality in Vietnamese, in order to build a complete temporal system, based on deep linguistic and socio-cultural analysis.

In the following, we outline the framework in which we wish to contribute in one hand to the resolution of these impediments and, in the other hand, to enrich the annotation in order to include a temporal reasoning system. In the next section, we give a general survey of our multilevel annotations description. We develop the interest - particularly for a temporal reasoning task - of choosing the S[et]-languages framework [7] for the annotation of temporal relations provided by deictic calendrical expressions. Then we present the implementation part.

II. About annotations

Based on the presupposition that two expressions that refer to the same object differ on what they mean about the situation of the discourse (either aspectual, culturally, intentionally, and so on), we propose a multilevel annotation description in two parts: one for the reference and one for the situation of discourse (in large). We aim at providing a general framework of annotation taking into account the two parts of the description and the possibility of interconnection between other annotating frameworks such as TimeML and the temporal reasoning tool SLS developed by Irène Durand in labri, based on the Sylviane S. Schwer’s S-languages framework [8]. The first step, in development, concerns the calendrical unit “day”, which is the basic and more complex one, due to the fact that in one hand, this is the canonical measure of time, and or the other hand, the day unit is based on the most fundamental light/night alternative which still determines (in some extend) activities and rests of human beings.

The temporal reference part: This part can be viewed as the classification of all recognizing calendrical expressions containing these three words according to their denotation. For instance, the three Viet. expressions *hôm nay, bữa hôm nay, ngày hôm nay* denote the same deictic relationship: they relate some processes inside a unit-day-period that contains the moment of speech (coded S). *bữa nay* and *ngày nay* denote another deictic relationship: they relate some processes inside a period (to be further specified) that contains the moment of speech. The Fr. *Aujourd’hui* subsumes the five preceding Vietnamese expressions, and

could be described in term of a period. In order to deal also with temporal reasoning, these relationships are encoded in terms of S-languages, which can be viewed as an algebraic formal language that generalizes the reichenbachian way of coding temporal relations expressed by tense [9]. Let us go on with calendrical deictic expressions, that is relative to the moment of speech.

Each calendar unit is associated to a letter (ex. **d** for day, **y** for year). An occurrence of a letter depicts a bound (the | in Figure 1) between two successive durations of the unit.

------|-----|---E'---|---S--E-|------|------|---E"----|

Figure 1

For instance, *today* is described as **+d[S⊗?]d+.** In this expression, **"?"** is a letter that signals something is said to occur inside the duration defined by **dd** and **"[S⊗?]"** says that the expression gives no information that allow to order **S** and **?**. The exact temporal relation between S and E is given either by the cotext or by the situation of discourse.

+d and **d+** signify that **d** is iterable both backward and forward.

+d[S⊗E]d+ encodes the fact that event E occurred/occurs/will occur today. *Yesterday* is described as **+d?dSd+,** *tomorrow* is described as **+dSd?d+,** *ten days ago* is described as **+d?d^{10}Sd+,** *this year* is described as

+y[S⊗?]y+, *ten years ago* is described as **+y?y^{10}Sy+.**

Let the sentence (5) be an illustration of Figure1,

(5) *Yesterday I saw Eve (E'), today I will see Jean*
(E) and in three days Zoé
(E").

its S-expression is the result of the Join operation

+dE'dSd+, +d[S⊗E]d+, SE[1],+dSd^{3}E"d+, that is:

+dE'dSEd^{3}E"d, which is depicted Figure 1, where | stands for **d.**

SLS [7] is a software that computes the Join of any set of S-expressions. An interface between Nooj and SLS is intended.

Based upon the S-language annotation, it is trivial to see that **+d?dSd+** and **+y?dSy+** can be subsumed by the schema **+u?uSu+,** that describes that something is said on the unit preceding the deictic anchoring unit encoded by **u.** This allows a hierarchical structure of expressions, not based on the kind of unit, but only on any temporal partitioning and the proximity to the deictic anchor. For instance, **u** can encoded any saillant type of events such that a sunami occurrence: in that case **+u?uSu+** codes a process that occurred between the the sumami that preceded the last one and the last one. This is also the good way to encode the *Fr* deictic *hier/aujourd'hui/demain,* that is **+u?uSu+, +uS⊗?u+ , +uSu?u+** with **u=d** as in (1) and u= change of period, to be determined as political phase of a state in (2).

The contextual part: this is the socio-linguistic analysis part. The discrimination will be made between each expression of the same class. This requires a deep understanding of the language, in all its aspects. Vietnamese belongs to the group of isolating, tonal and monosyllabic languages. Its word forms are invariant, there is no verbal tense. Hence, all grammatical relations are expressed by word order and/or function words and/or tonal and prosody device. Temporality is strongly affected by these characteristics. All linguistic domains have to be questioned.

• *Lexically*: Northern Vietnamese uses *ngày* while Southerners prefer *bữa,*

• *Syntactically*: the position of the temporal adverbial in a question/answer indicates the future or the past : *Khi nào anh về* (future) vs *Anh về khi nào* (past),

• *Semantically*: tone has a semantic value and can be used to describe proximity/remotness: *hôm kia/kià/…* [10],

[1] This co-textual constraint is given by the future tense.

• *Pragmatically*: cultural and social aspects linked to calendrical systems and time perception : *âm lịch / dương* .

Beyond these classic ones, the following aspects can't be neglected. Vietnamese NLPA studies such as [11, 12] testify of their importance.

• *Tonaly*: specific tonal transformation in regional dialect describing a temporal proximity: in Southern dialect, "*hổm*" is equivalent to "*hôm ấy*".

• *Prosody*: (i) the use of reduplication can be used for vagueness: *chiều chiều "epoq epoq" for* "near this afternoon" or for iteration *ngày ngày* "jour jour" for "each day"; (ii) the preference for the bisyllabilism in numbering, which obliges to add *mùng* before numbers 1 to 10.

The Vietnamese study of temporality is generally centered on phenomena related to the predicate, particularly tense and aspect and adverbs. Temporal indicators that lie outside this category, most of which involve calendrical terms, have been quite ignored [8],[10], or [11,12].

For modeling the temporal elements in Vietnamese we proceed as a second language learner who want to understand how to construct a sentence with temporal marks. So we look in many grammatical [13,17] and pedagogical books [14, 15, 19], dictionaries [22,23] and lexical [20] of Vietnamese for non-native speakers. Despite their useful analysis, the complements of tense and temporal frames are not completely mentioned and well studied. These works often focus on communication skills without providing an "immediate constituent analysis", which would be helpful for our analysis. For instance, in [5], "*today*" is translated by five expressions:

(6) *bữa nay, hôm nay, ngày nay, bữa hôm nay, ngày hôm nay.*

But there is no explanation about the meaning and the syntactical combinatory possibilities of *hôm, ngày,* and *bữa*. Actually the automatic translation systems have difficulty in differentiating the constructions with «*hôm* » and « *ngày* », expecially when the co(n)text can't be found in their table-phrases, as shown in example (4).

Some explanations can be found in [21], where is said that *bữa* is used in Saigon, "as one of a series, or time when, primarily in present or future", *hôm* "as time when, primarily in present or past", and *ngày* "as one of a series, or time when, in future", *Today* is translated by *hôm nay* in Hanoi and by *bữa nay* in Saigon. Nowadays, it seems that *hôm nay* is used in both area to mean the calendrical present day, *bữa hôm nay* only in Saigon and *ngày hôm nay* in Hanoi. *Bữa hôm nay* and *ngày hôm nay* are used for today/nowadays. Following [24] and [25], we can argue that the trisyllabic expressions maintain the calendrical today inside its calendrical serie, that is, in constrast with yersterday and tomorrow. These explanations provide a reason to have the five different expressions given in (6) for translating "today" with a day-unit term and the way to differenciate them?

III. A FIRST GRAMMAR FOR *HÔM* AND ITS IMPLEMENTATION

This paper constitutes a first work about Vietnamese temporality detection and formalization. It consists not only in providing results of textual treatment but also to test the robustness of detecting temporality structures in texts. Our goal is to provide a set of annoted grammars for each calendrical noun. We have begun with day-unit terms and *hôm* was choosen as a test, because it is in opposition with the two others.

A 1516 articles corpora was build from the «Thế Giới » category of the Vietnamese website VnExpress.net. We detected then 616 occurrences of «*hôm*» by using the NooJ's frequency functionality.

NooJ[3] is a language environment developed by Max Silberztein [3] with a wide open community having developped a set of linguistic tools for more than twenty languages. As we mentioned it above, the Viet4NooJ (V4N) project refers to the

linguistic resources created specifically for data treatment in Vietnamese [4], [5], [6].

A few reasons explain why NooJ has been chosen for our work relative to its specificities:

Flexibility: NooJ's functionalities are is final end-user oriented plateform. The interface was elaborated to facilitate the creation of linguistic resources for non NLP specialist.

Granularity: NooJ uses three particularly efficient computational devices to process texts: (i) Finite-State Automata (FSA) to locate specific expression in corpora, (ii) Recursive Transition Networks (RTN), a grammar coomposed by a set of sub-graphs and (iii) Enhanced Recursive Transition Networks (ERTNs), a set of RTN performing textual operations through the use of variables, producing specific outputs.

Interoperability: ERTNs' results (i.e. output / tags) can be re-used in external software. Thus, NooJ would be able to play the role of intermediary phase between heterogeneous copora and computational calculation of time relations with [8] as final results.

Multimodal tagging process: the system is able to produce specific tags (i.e. created by the user) or xml tags as well. This permits to produce TimeML tags as previously mentioned.

Studies about temporal expressions recognition can be schematically described under two main logics : the first one concerns a pattern matching rules approach and the second one, through supervised learning methods [26]. The semantic based models like Timex2, Chronos or TimeML are the most common systems used by the researchers community working on temporal expression. These models annote temporal expression with two steps: (i) marking a temporal expression in a document, and (ii) identifying
the time value that the expression designates. Studies are, until now, focussed on English and european languages such French, German, Italian, slovenian and serbian. Since 2005, studies on Chinese and Korean begun to be developping essentialy to normalize temporal expression and provide tests for both translation and alignement [27, 28].

Concerning more specifically *hôm*, FSA and RTN were organized to produce a specific ERTNs aiming at detecting the maximal length expression with *hôm* as core element (Fig. 2). Morphology of Vietnamese language is monosyllabic, every morphem being composed by only one vowel or vocalic component. Graph and subgraph of our ERTNs follow a semantic logic instead of a lexematic one. For example, the expression "*hôm thứ ba*" [*hôm* : day, *thứ* : ordinal marker, *ba* : numeral "3" -> *thứ ba* : tuesday] with 3 morphological elements will be tagged with only two hierarchical levels : <hôm,N0><thứ ba,N+1>.

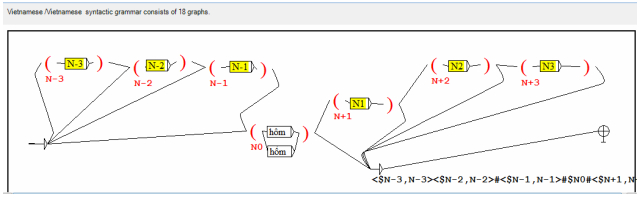

Figure 1 : The ERTNs for *Hôm*

This ERTNs is constituted by a set of ten subgraphs organized around the "*hôm*" nucleus and its regional variant "*hổm*". The ERTNs is actually able to recognize temporal expression with hôm with a hierachical positioning levels of N-3 -> /…/ N0 (hôm) /->/…/ N+3

The Table 1 gives an example of a relevant expression extracted by V4N:

N-3	**N-2**	**N-1**	**N0**	**N+1**
Rất	*sớm*	*sáng*	*Hôm*	*thứ ba*

Very early in the morning on Tuesday

Table 1 : Hierarchical levels of Hôm : N-3/ 0/N+1

IV. Results and Discussion

Application of the "*hôm*" ERTNs over the corpora provides good results since the return rate approximates 99 %. However, this result has to be relativized : the test was only conducted on an electronic press sample. We assume that results would have been different with a literary sample. This new test will be son explored to prove the robustness of our grammar and to

enhance it by adding new lexems to the different levels of the grammar.
Beside this, our work can be considered as an opening to some new perspectives on linguistics and NLP for Vietnamese since, as far as the state of art has shown it, till now, almost all of the temporality studies has been focused on temporal lexicography level rather than on detection of temporal framework. Our approach is also bearing a few secondary emerging topics as enhancement of praxematic process, psycho-cognitive studies, didactic process for non native learners to better understand language structuration and reasoning logic, cross-cultural information retrieval, etc.

V. REFERENCES

[1] J. Pustejovsky, J. Littman, R. Saurí, and M. Verhagen, "TimeBank 1. 2 Documentation," Event London, no. April, pp. 6-11, 2006.
[2] T. ngoc diep Do, "Extraction de corpus parallèle pour la traduction automatique depuis et vers une langue peu dotée," CNRS:UMR5217 – INRIA – Université Pierre Mendès-France - Grenoble II – Université Joseph Fourier - Grenoble I – Institut Polytechnique de Grenoble -, 2011.
[3] M. Silberztein, "NooJ: a linguistic annotation system for corpus processing," 2005, ," in *Proceedings of HLT/EMNLP on Interactive Demonstrations* -, 2005, pp. 10-11.
[4] P. Lambert and M. Fournié, "A Vietnamese module for NooJ: Modelization, realization and perspectives," 2008.
[5] P. Lambert, M. Fournié, and O. Ho Dinh, "VIET4Nooj A Vietnamese module for Nooj," in NooJ conference 2010, 2010.
[6] N. Boffo and O. Ho Dinh, "Automatic Processing of Temporality for VIET4Nooj," in NooJ 2010 Conference in Komotini (GR), 2010.
[7] S. R. Schwer, "Représentation mathématique du temps : après Reichenbach," Tranel, no. 45, pp. 167-186, 2006.
[8] I. A. Durand and S. R. Schwer, "A Tool for Reasoning about Qualitative Temporal Information: the Theory of S-languages with a Lisp Implementation," Journal Of Universal Computer Science, 14, pp. 3282-3306, 2008.
[9] H. Reichenbach, *Elements of Symbolic Logic*. Free Press, New York, 1947.
[10] P. P. Nguyễn Questions de linguistique vietnamienne : les classificateurs et les déictiques /. Paris:Presses de l'Ecole française d'Extrême-Orient,1995.
[11] D. K. Mac, E. Castelli, and V. Auberge, "Modeling the prosody of Vietnamese attitudes for expressive speech synthesis," in International workshop on Spoken Langages Technologies for Under-resourced languages.
[12] L. H. Phuong, N. T. M. Huyen And A. Roussanaly. Finite-state description of Vietnamese reduplication IN Proceedings of the 7th Workshop on Asian Language Resources 2009
[13] K. T. Nguyễn, *Nghiên cứu ngữ pháp tiếng Việt*. Hà Nội: NXB Giáo dục, 1997.
[14] Diệp Quang Ban, Ngữ pháp tiếng Việt [Grammaire vietnamienne]. Hà Nội: NXB Giáo dục, 2005, p. 671.
[15] V. L. Lê, Le parler vietnamien. Saigon: 1960, p. 281.
[16] H. Goto, Y. Hasegawa, and M. Tanaka, "Efficient scheduling focusing on the duality of MPL representation," in IEEE Symposium on Computational Intelligence in Scheduling 2007 SCIS07, 2007, vol. 053920, pp. 57–64.
[17] Đoàn Thiện Thuật, Nguyễn Khánh Hà, Phạm Như Quỳnh, A concise vietnamese grammar for non-native speakers, VNU, Hanoi, Thế Giới Publishers, 2006.
[18] Vũ Văn Thi, Tiếng việt cơ sở, Nhà Xuất Bản Đại Học Quốc Gia, Hà Nội, 2008.
[19] Ngyễn Việt Hương, Tiếng việt nâng cao dành cho ngươi ngoài (quyển 1), Nhà Xuất Bản Đại Học Quốc Gia Hà Nội, 2010.
[20] Nguyễn Anh Quế, Tiếng việt cho ngươi nước ngoài, Nhà Xuất Bản Văn Hóa, Thông Tin, 2000.
[21] Lê Khả Kế, Từ điển Pháp Việt, Nhá xuất bản Thành Phố Hồ Chí Minh, 2001.
[22] Từ điển Tiếng Việt, Vietlex, trung tâm từ điển học, Nhà xuất bản Đà Nẵng, 2009.
[23] Từ điển Việt Pháp, Viện Khoa học xã hội Việt Nam, Nhà xuất bản văn hóa Sài Gòn, 2001.
[24] L. C. Thompson, "A Vietnamese Grammar," *MonKhmer Studies A Journal of Southeast Asian Linguistics and Languages*, vol. 13–14, no. 4, p. xxi+386, 1965.
[25] S. R. Schwer, Au jour d'aujourd'hui *Cahiers AFLS* 17.2 p. 3-31, 2012.
[26] Sanampudi, S. K., & Kumari, G. V. (2010). Temporal Reasoning in Natural Language Processing: A Survey. International Journal of Computer Applications, 1(4), 68-72.
[27] Lecuit, E., Maurel, D., Vitas, D., & Krstev, C. (2006). Temporal Expressions□: Comparisons in a Multilingual Corpus. October, 531-535.
[28] Wu, M., Li, W., Chen, Q., & Lu, Q. (2005). Normalizing Chinese Temporal Expressions with Multi-label Classification. Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering (pp. 318-323).