Toward Time Universals Identification in NooJ : A New Model of Time Recognition in Vietnamese

Sylviane R. SCHWER (Université Paris 13, Sorbonne Paris Cité, LIPN, CNRS UMR 7030).

Nicolas BOFFO (Praxiling/Université Montpellier 3 et MICA/Institut Polytechnique de Hanoi).

Philippe LAMBERT (Groupe de Recherche sur le Vietnam Contemporain, GReVIC, Vinalor)

Plan

I. Objectives

II. Linguistic positions and orientations

III. Annotation with Automatic Processing of Temporality for Vietnamese Component APTVC : the case the day-expressions constructed with Hôm and Ngay.

IV. Construction of Universal Temporal Classes

I. Objectives

- The semantic annotation of temporal expressions with Automatic processing of temporality for vietnamese component APTVC V 0.2.0
- The establishment of Universal Temporal Classes (UTC)
- The implementation of this system in NOOJ to compute temporal relations between processes

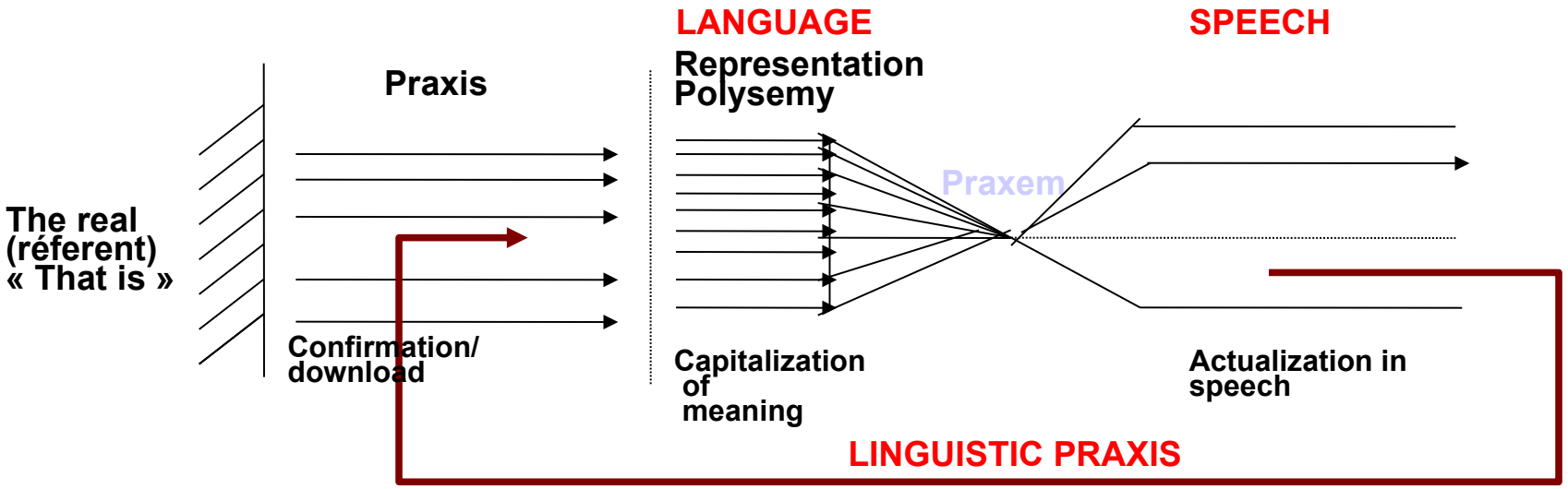
II. Linguistic positions and orientations

Our positions and orientation in linguistics: Praxematic

The main idea praxematic is that humans base their linguistic representations on their praxis (experience).

The aim : the production of global meaning from each pieces of informations (lexical, syntactical, semantical, textual, pragmatical)

A strongly impediment : polysemy



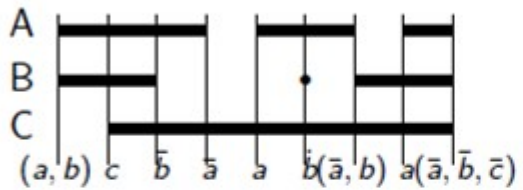
Our framework

- **Linguistics point of view : each temporal marker of a text produces some kind of instructions that had to be joined with the others to produce the temporal meaning (Jacques Bres).**
- **Vietnamese temporality : no verbal tense, aspect and tense markers, temporal frame-adverbial phrases, syntactical position …**
- **TAL point of view : the internal structure of temporal expressions must be investigated in order to construct grammars that can capture general features and be used in automatic processing system.**
  - **Detection : by temporal frames (BOFFO, 2010).**
  - **Annotation : anything which help to understand the exact meaning(s ?) (grammars and corpus analysis) and to allow temporal reasoning .**

Temporal reasoning : based on S-languages theory.

The choice of one annotation for the temporal relations description : S-langages

S-langages
Exemple



son S-mot $(a\bar{a})^3$

son S-mot $b\bar{b}\dot{b}b\bar{b}$

son S-mot $c\bar{c}$

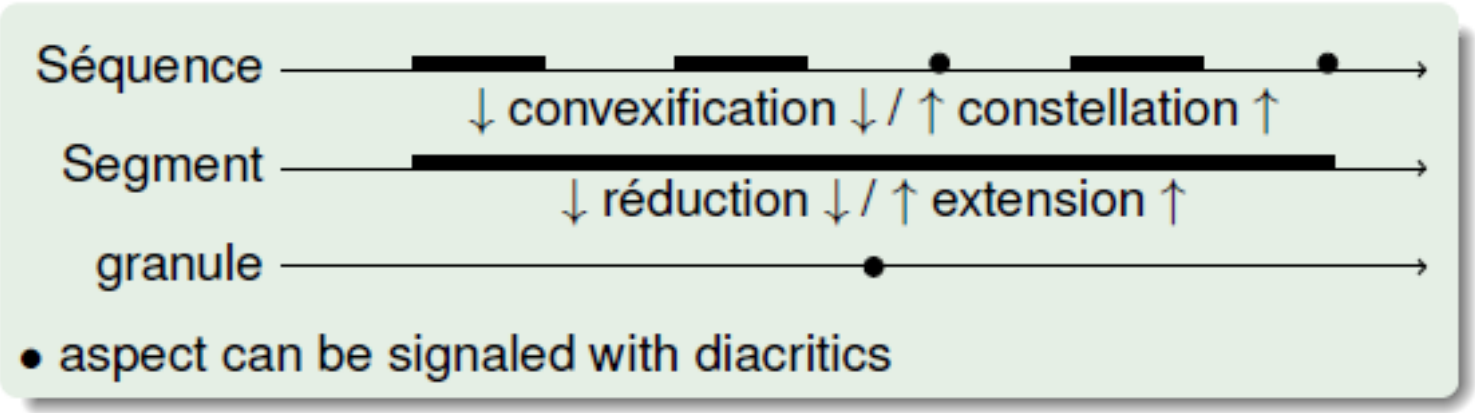
Le S-mot encodant la relation

- les entités : A, B, C
- l'alphabet : $\mathcal{A} = \{a, b, c\}$, le vecteur de Parikh : (6,5,2)
- le S-alphabet :
  $\widehat{\mathcal{A}} = \{(a), (b), (c), (a, b), (b, c), (a, c), (a, b, c)\}$
- description des objets : $(a)^6$, $(b)^5$,$(c)(c)$
- description de la relation: (a,b)(c)(b)(a)(a)(b)(a,b)(a)(a,b,c)
- Le S-univers :
  $\mathcal{L}(6, 5, 2) = \{u \in \widehat{\mathcal{A}}^* || u|_a = 6, |u|_b = 5, |u|_c = 2\}$

Granule

aspectuality and the notion of granule

- some process can be viewed as punctual, or durative or as a series, ...
- The *granule* : a durative indivisible point (specious point)
  - to modelize the neutral aspect and to reduce the complexity of computations.
  - Two granules can be contiguous, that is forbidden for two mathematical points.



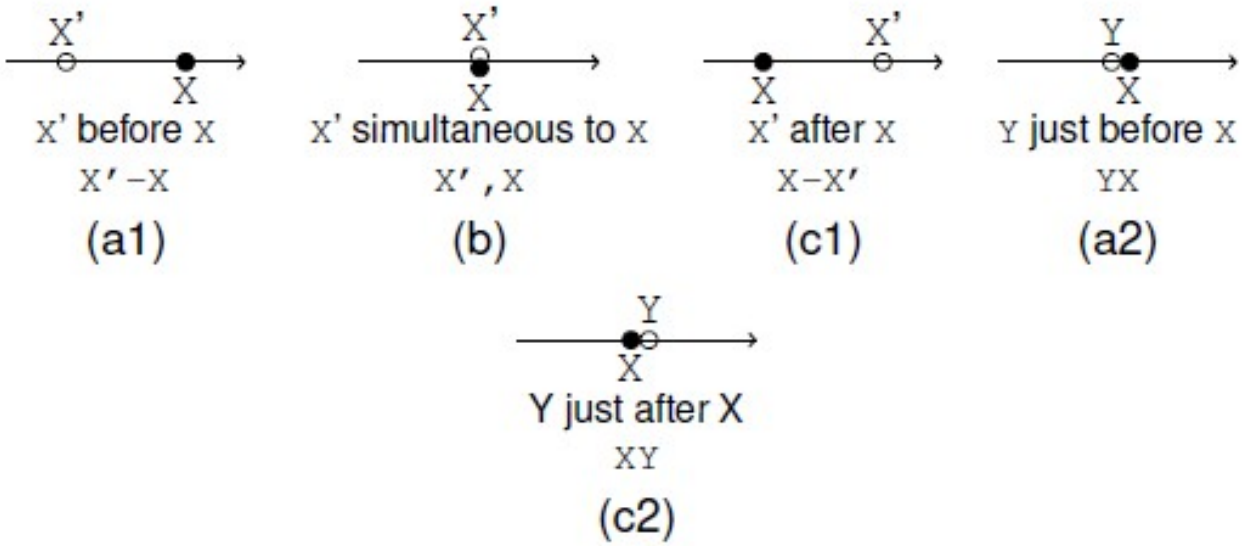
- aspect can be signaled with diacritics

instant : x	ponctual : $\dot{x}$	period : $\bar{x}$	series : $\ddot{\bar{x}}$

- temporal relations motivate the choice of representations

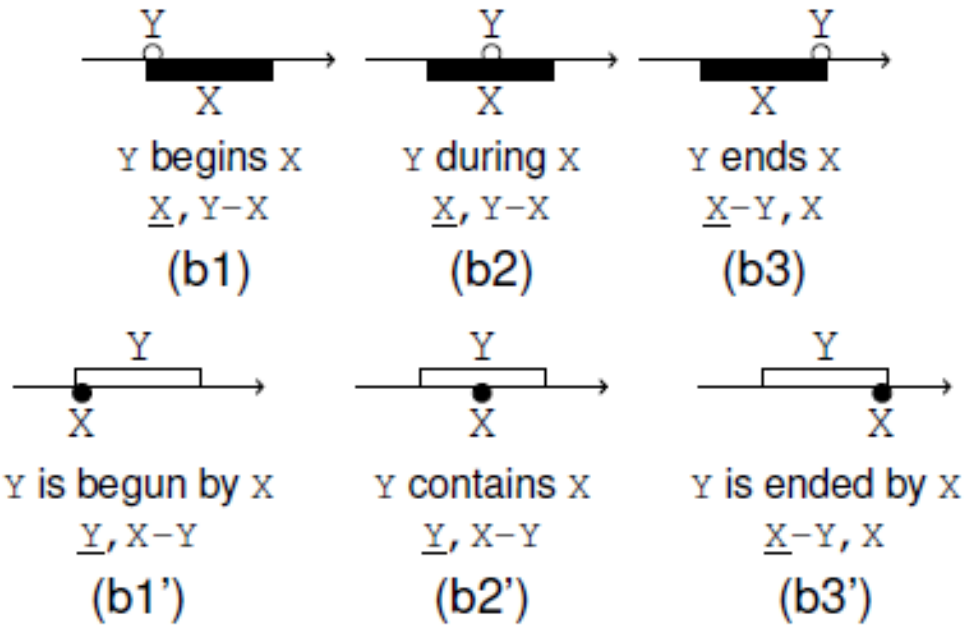
Temporal Relations between 2 granules

relative situations between two granules precedency and contiguity

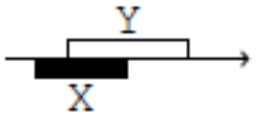


Temporal Relations betwen a granule and a period

inclusions relation : a granule in a period



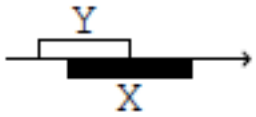
Temporal Relations beetwen two periods : never described in one information



Y begins during X and ends after X

X–Y–X–Y

(bb)



Y begins before X and ends during X

Y–X–Y–X

(bb')

- Simultaneousness is in fact the non before/after relation : all the (b) can be together rewrited as

$$XoY = \textit{neither } (X \textit{ before } Y) \textit{ nor } (Y \textit{ before } X)$$

III. Annotation with Automatic Processing of Temporality for Vietnamese Component APTVC : the case the day-expressions constructed with Hôm and Ngay.

Hôm / Ngày

Hôm and *Ngày* both mean « day » but not in the same way.

H*ôm and Ngày are two nucleus whose build temporal expressions.*

Actually the automatic translation systems have difficulty differencing the constructions with «*hôm* » and « *ngày* », expecially when the co(n)text can't be found in their table-phrase as Google Translation System.

Exemple with Google translation

Fr. Source:

« Aujourd'hui *nous sommes dans une situation économique difficile."* into the *Viet.*

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 period. One has to choose between ***ngày nay*** or ***hiện tại***.

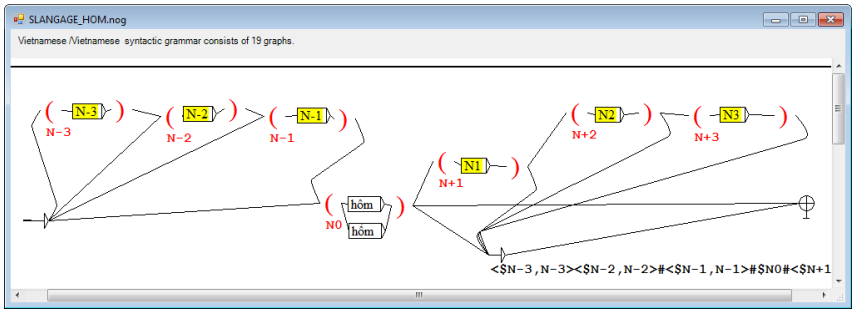
Tables construction of *hôm* & *ngày* temporal expressions

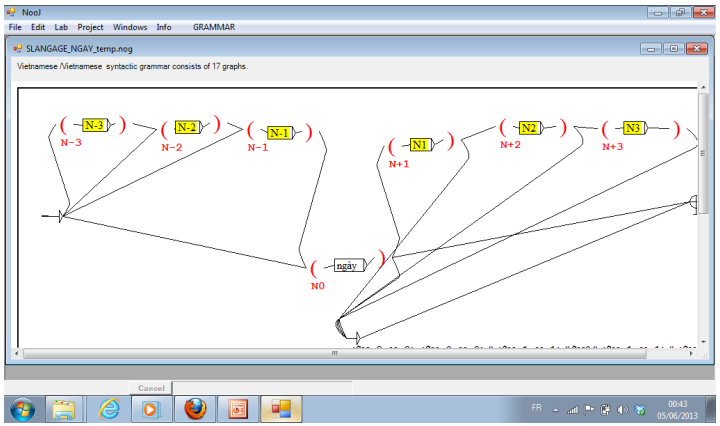
To understand the production of global meaning and also to prepare NOOJ's graphs building , we extract most of temporal expressions constructed with *hôm* & *ngày* nucleus on newspapers corpus (1868 articles of various press) and vietnamese grammar books :

- Manually by linguistic expertise,
- and automaticly with coocurence and mutual information extraction tools: PERL script and T*rameur* tools from Serge Fleury (Paris 3 TAL) and NOOJ's occurences.

N-3	**N-2**	**N-1**	**Nucleus**	**N+1**	**N+2**	**N+3**

Hôm & Ngày graph s





An example of a relevant expression extracted by Viet4Nooj :

N-3 N-2 N-1 N0 N+1

Rất sớm sáng Hôm thứ ba

Very / early / in the morning [DAY]/ on Tuesday

We detected 616 occurrences of «*hôm*» by using the NooJ's frequency functionality and our return rate approximates 99 % .

- Ngay is essentially use for the calendar expressions.

Ex: 2 | ngày | 3| tháng| 9 |năm| 2012

2 | day | 3| month| 9 |year| 2012

We detected arround 1400 occurrences of «Ngay» with NGAY graph.

IV. Construction of Universal Temporal Classes (UTC)

Formalization of UTC by an proposition of algorithm « *package UniversalTemporalClasses* »

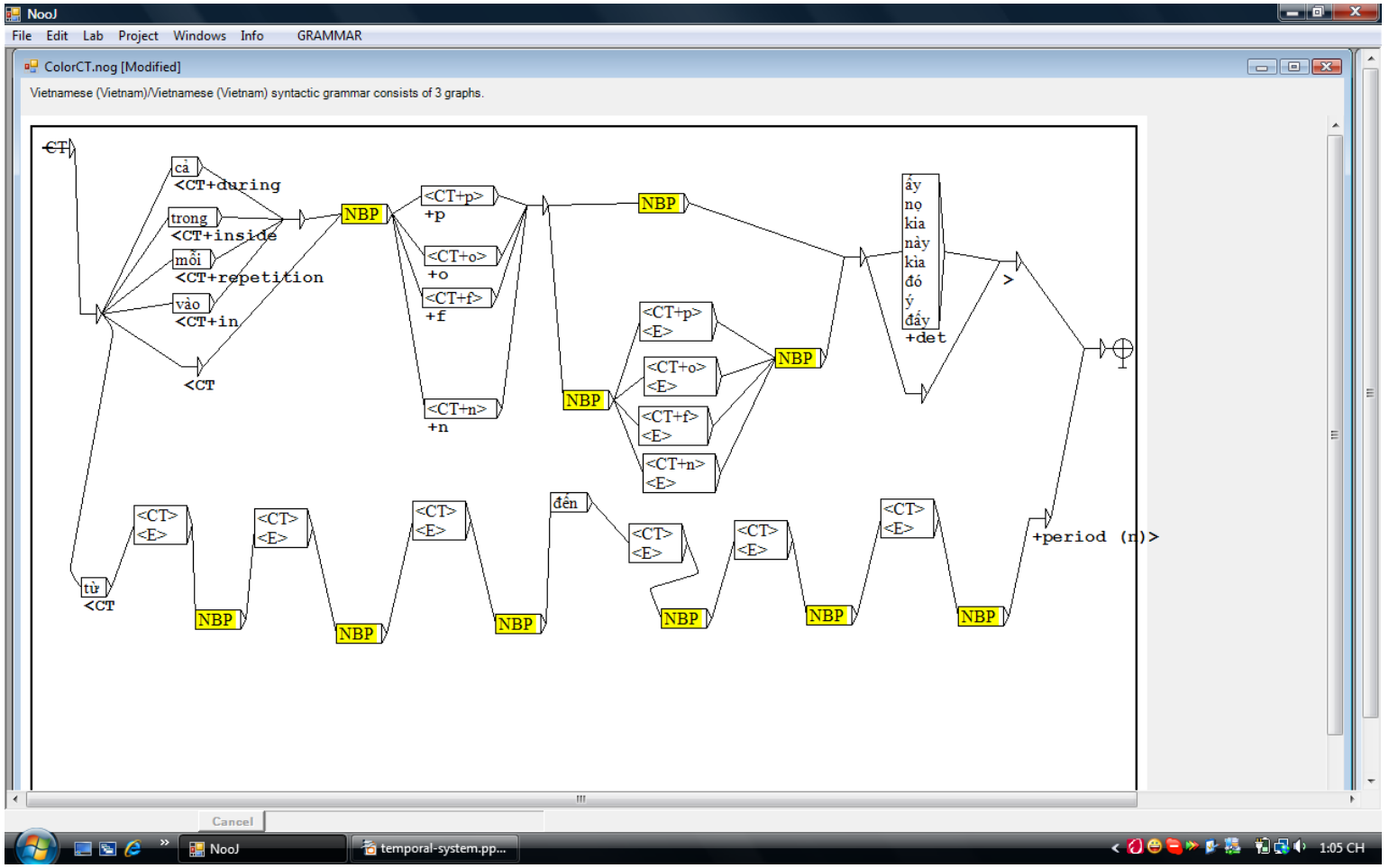
A first list of Temporal TAGS

(Nicolas BOFFO, 2010)

impediment = not exhaustive and complete

Time	Aspect	**TemporalFrameworksTags :**
<Past> **<Present>** **<Future>** **<Neutral>**	**<Accomplished >** **<In progress>** **<Iterative>** **<Terminative>** **<Imminent>**	**<CT+During>,** **<CT+Inside>,** **<CT+Repetition>** **<T+Period>** **<open>** **<close>**

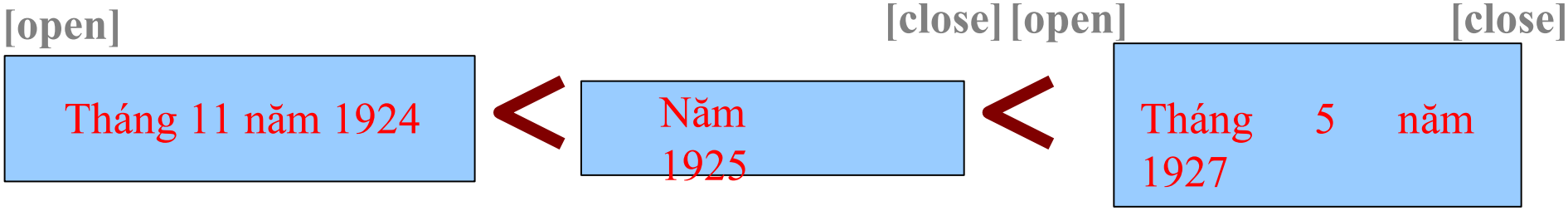
Firts implementation in NOOJ was for Temporal Framework* Graph



Temporal Frames in Biograhy of Ho Chi Minh

[open]Tháng 11 năm 1924, Nguyễn Ái Quốc về Quảng Châu (Trung Quốc) .... Năm 1925, Người **thành lập** Hội Việt Nam Cách mạng Thanh niên .....**[close]**

[open]Tháng 5 năm 1927, Nguyễn Ái Quốc rời Quảng Châu **đi** Mátxcơva (Liên Xô) ,sau đó **đi** Berlin (Đức), **đi** Bruxell (Bỉ) **tham dự** phiên họp mở rộng của Đại hội đồng Liên đoàn chống chiến tranh đế quốc, sau đó **đi** Ý và từ đây về Châu Á.**[close]**



We are creating an
Universal Temporality Package

We wish to implement it in NOOJ

The attributes of our UTC

Based on TimeML and our linguistic approaches:

tid ::= ID

pos ::= 'ADJECTIVE' | 'NOUN' | 'VERB' | 'PREPOSITION' | 'OTHER‘

tense ::= 'FUTURE' | 'INFINITIVE' | 'PAST' | 'PASTPART' | 'PRESENT' | 'PRESPART' | 'NONE'

PraxematicTenseInstruction::[+PAST] ' | ‘ [NEUTRAL] ' | ‘ [+FUTURE]

aspect ::= 'PROGRESSIVE' | 'PERFECTIVE' | 'PERFECTIVE_PROGRESSIVE' | 'NONE' | ‘ **ACCOMPLISHED** ' | ‘**ITERATIVE** ' | ‘**TERMINATIVE** ' | ‘ **IMMINENT**

PraxematicAspectInstruction::= [+TENSION] ' | ‘ [+EXTENSION] ' | ‘ [-INCIDENCE] ' | ‘ [-INCIDENCE][BI-EXTENSION]

TemporalFrameworkValue ::= <CT+During> ' | ‘ <CT+Inside> ' | ‘ <CT+Repetition> ' | ‘ <T+Period> **(to review)**

Phrase ::=<open> ' | ‘ <close>

SyntacticPositionFromNucleus::=[N^0] ' | ‘ [N^{n+1}] ' | ‘[N^{n-1}]

UTC’s attributes

relType ::= 'BEFORE' | 'AFTER' | 'INCLUDES' | 'IS_INCLUDED' | 'DURING' | 'SIMULTANEOUS' | 'IAFTER' | 'IBEFORE' | 'IDENTITY' | 'BEGINS' | 'ENDS' | 'BEGUN_BY' | 'ENDED_BY' | 'DURING_INV'

relType ::= 'MODAL' | 'EVIDENTIAL' | 'NEG_EVIDENTIAL' | 'FACTIVE' | 'COUNTER_FACTIVE' | 'CONDITIONAL'

relType ::= 'INITIATES' | 'CULMINATES' | 'TERMINATES' | 'CONTINUES' | 'REINITIATES'

mod ::= 'BEFORE' | 'AFTER' | 'ON_OR_BEFORE' | 'ON_OR_AFTER' |'LESS_THAN' | 'MORE_THAN' | 'EQUAL_OR_LESS' | 'EQUAL_OR_MORE' | 'START' | 'MID' | 'END' | 'APPROX'

UTC’s attributes (end)

cardinality ::= CDATA

polarity ::= 'NEG' | 'POS' {default, if absent, is 'POS'}

type ::= 'DATE' | 'TIME' | 'DURATION' | 'SET‘

beginPoint ::= IDREF
endPoint ::= IDREF

quant ::= CDATA

freq ::= Duration

value ::= Duration | Date | Time | WeekDate | WeekTime | Season | PartOfYear | PaPrFu

comment ::= CDATA

An exemple UTC analysis

Ex: **|Hôm ||qua|| tôi| đang |xem| TV| thì |điện thoại |reo.**

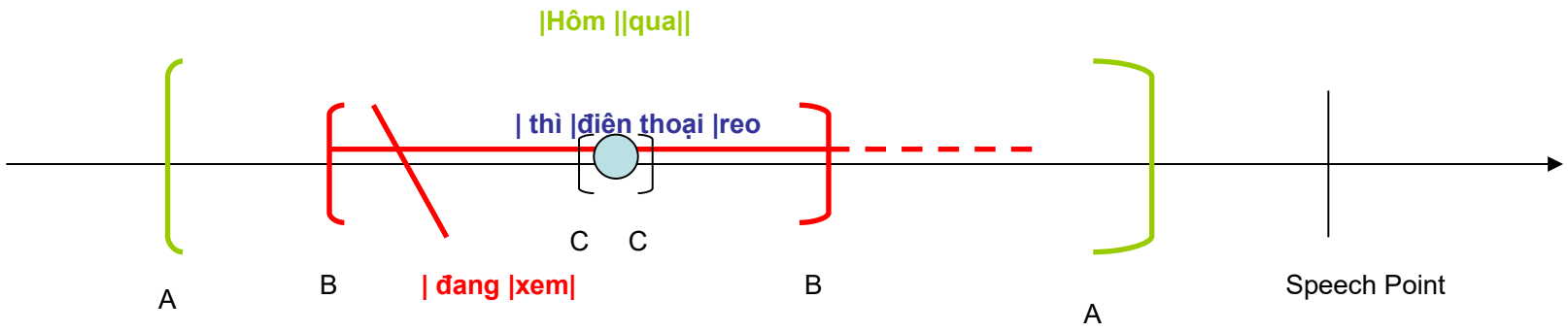
|Day ||to be gone|| I |aspectual mark |watch| TV| then |telephon |ring.

Annotation:

<open> | Hôm<N0> | | qua <N+1><ACCOMPLISHED> | <+PAST> |

| tôi | đang<-INCIDENCE><PROGRESSIVE> | xem<V> | TV |
thì<MID> | điện thoại | reo<+INCIDENCE><ACCOMPLISHED> |
<CLOSE>.

Graphical representation:



Thank you
for your attention!

Bibliography Elements :

- J. Pustejovsky, J. Littman, R. Saurí, and M. Verhagen, “TimeBank 1. 2 Documentation,” Event London, no. April, pp. 6-11, 2006.
- 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.
- M. Silberztein, “NooJ: a linguistic annotation system for corpus processing,” 2005, ,” in Proceedings of HLT/EMNLP on Interactive Demonstrations -, 2005, pp. 10-11.
- P. Lambert, S. Schwer, N. Boffo, “A new model of Time Expressions Detection and Annotation in Vietnamese : The *hôm* case” in International Conference on Asian Language Processing, 2012, Hanoi.
- M. Charolles, L’encadrement du dicours : Univers, Champs, Domaines et Espaces. Paru en 1997 dans Cahier de Recherche Linguistique, LAN DISCO, URA-CNRS 1035 Université Nancy 2, n° 6, 1-73
- Charolles M. Les cadres de discours et leurs frontières». In D.Delomier & M - A. Morel eds. Frontières : du linguistique au sémiotique. Lambert - Lucas, Limoges, 2009 ,143 -162.

- [P. Lambert and M. Fournié, “A Vietnamese module for NooJ: Modelization, realization and perspectives,” 2008.
- P. Lambert, M. Fournié, and O. Ho Dinh, “VIET4Nooj A Vietnamese module for Nooj,” in NooJ conference 2010, 2010.
- N. Boffo and O. Ho Dinh, “Automatic Processing of Temporality for VIET4Nooj,” in NooJ 2010 Conference in Komotini (GR), 2010.
- S. R. Schwer, “Représentation mathématique du temps : après Reichenbach,” Tranel, no. 45, pp. 167-186, 2006.
- 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.
- [H. Reichenbach, Elements of Symbolic Logic. Free Press, New York, 1947.
- [P. P. Nguyên Questions de linguistique vietnamienne : les classificateurs et les déictiques /. Paris:Presses de l’Ecole française d'Extrême-Orient,1995.