「언어연구」, 제25권 제3호(2008), 39-63

프랑스어 시간개체 표현의 정규화를 위한 기초연구*

윤애선
(부산대학교)

Yoon, Aesun. 2008. Normalization of Temporal Expressions in French: A Preliminary Study. *Linguistic Research* 25(3), 39-63. The primary purpose of this study is to examine some characteristics of temporal expressions in French to normalize their annotation with the framework of ISO-TimeML (2008). Little work has been done on the automatic searching of such expressions and their semantic annotation other than recent work related to ISO-TimeML (2008) and its precursor TimeML (2006). Gross (2002) proposes automata-based searching methods for two types of TIME and DATE referring expressions in French, while its successor Bittar (2008) treats two more types, DURATION and SET, but with little description of their components. This paper examines each component of these four types and proposes the templates that describe searching automata for a variety of temporal expressions associated with their possible attribute-values properly specified. **(Pusan National University)**

Keywords normalization, temporal expression in French, ISO-TimeML, annotation, TIMEX3

1. 들어가기

시간은 인식의 주체인 인간의 경험을 구성하는 가장 중요한 요소 중 하나이며, 흔히 과거에서 미래로 나아가는 방향성을 가진 선으로 표상된다. 따라서 실제 경험 여부를 떠나 인간이 인식할 수 있는 모든 사건은 선형의 시간축(linear time axis) 위에 위치하게 되면, 각각의 사건은 그 시간선 상의 상대적 위치에 따라 전후 관계 등을 갖게 된다. 이러한 인식이나 경험을 언어로 기술하면, 다양한 종류의 시간개체 표현(temporal expression)을 통해 나타날 수 있다.

몇 가지 시간개체 표현이 나타난 간단한 예를 들면 다음과 같다.

* 이 논문은 2007년도 부산대학교 인문사회연구기금의 지원을 받았음. 이 논문의 초고를 꼼꼼하게 읽고 부족한 부분을 지적해 주신 이기용 선생님, 손현정 선생님, 이은령 선생님과 심사자께, 지면을 빌어 심심한 사의를 표한다. 그래도 여전히 부족한 부분은 저자의 몫이다.

(1) 최익현:
14세 때부터 이항로 문하에서 공부했다. 철종 6년 정시 문과에 급제하였으나, 고종 5년 경복궁 중건의 중지 등을 상소하여 관직을 삭탈당했다. 1906년 유배지인 대마도에서 사망하였다.

(2) a. 최익현 (단기 4166년 - 단기 4239년)
b. 철종은 1849년 19살의 나이로 즉위하였다.
c. 병인양요 발발 (단기 4199년, 고종3년)

(1)은 약 100년 전 인물인 면암 최익현을 소개하는 글이다. 이 짧은 글 안에 시간개체 표현[1]이 '14세 때, 철종 6년, 고종 5년, 1906년' 4개가 나타난다. 역사적 지식이 없으면 이 표현들이 가리키는 때를 전혀 알 수 없으며, 이러한 인물전은 일반적으로 먼저 일어난 것을 먼저 쓴다는 상식을 빼면 그것들이 가리키는 전후 관계도 알 수 없다. 예를 들어 '철종이 고종에 앞선다.'라는 기초적인 역사 상식이 있더라도 '최익현이 14세 되던 해', '철종 6년', '고종 5년'과 1906년을 동일한 시간선 상에 놓고 비교하기 어렵다. (2a-c)에 나타난 시간개체 표현도 마찬가지다. 따라서 위 4개의 예문만으로는 "최익현이 몇 세에 사망했는가?", "병인양요 발발 당시 최익현은 몇 세였나?", "최익현이 태어난 해는 서기 몇 년인가?", "최익현은 몇 세에 과거에 급제했는가?", "병인양요가 일어난 해는 서기 몇 년인가?", "철종이 즉위한 해는 단기 몇 년인가?"와 같은 매우 단순한 단답형 질문에 즉시 대답하기 어렵다.

4개의 예문에서 나타난 지식을 바탕으로 밑줄 친 시간개체 표현을 몇 가지 단순한 방식으로 정규화하면 [표 1]과 같다.[2] 사람도 다양한 사건의 시간적 관계를 파악하려면 서기든 단기든 재위 연도든 동일한 잣대를 가지고 이해하는 편이 훨씬 쉬울 것이다.

[표 1] (1)-(2)에 나타난 시간개체 표현의 단순한 정규화의 예

서기	(1833년)	(1847년)	**1849년**	(1855년)	(1866년)	(1868년)	**1906년**
단기	**4166년**	(4180년)	(4182년)	(4188년)	**4199년**	(4201년)	**4239년**
재위 연도			(철종1년)	(철종6년)	**고종3년**	(고종5년)	
외부 사건			철종즉위		병인양요		
철종의 나이			**19세**	(25세)			
면암의 나이		**14세**	(16세)	(22세)	(33세)	(35세)	(73세)
면암 관련 사건	출생	이항로 문하에서 수학 시작		문과급제		관직삭탈	

1) 이때 시간개체 표현은 이 논문 2절에서 설명하는 <TIMEX3>의 주석 대상이다.
2) [표 1]에서 진한 글씨체는 (1)-(2)에 명시적으로 나타난 시간개체 표현이고, 괄호 안의 연도는 이를 바탕으로 단순한 추론(시간계산)을 한 것이다.

이같이 같은 텍스트 안에 기술되는 여러 사건의 상대적 시간 관계나, 여러 텍스트에 편재한 관련 사건을 비교하기 위해서는, 시간을 동일한 잣대로 분절하고, 이를 동일한 방식으로 표상해야 할 수 있는 표준화(standardization)가 필요하다. 연도를 포함한 세밀한 시간성 표현을 국제적으로 표준화하려는 노력은 지난 10여 년간 ISO, MITRE, W3C 등의 OWL-Time, TIMEX, DAML-TIME, TimeML 등을 통해 꾸준히 제안되어 왔다. 초기에는 일정관리 전문가 시스템(scheduling agent)에서 필요한 매우 단순한 시간성 표현을 정규화하려고 했다 (Busmann & alii 1977). 형태와 통사 단계의 자연언어처리 기술이 발전하고, 의미 단계의 분석이 부분적으로나마 가능해지며 이를 통한 지식 추출과 추론이 가시화되면서, 텍스트에 나타난 다양한 시간 정보를 표상하려는 방향으로 발전하게 된다. 이와 같은 맥락에서 2007년부터 이러한 표준을 통합하여 ISO-TimeML을 만들고 있다.[3)]

ISO-TimeML은 MITRE 그룹의 TIDES 표준에서 제안한 <TIMEX2>를 발전시킨 TimeML를 근간으로 만들고 있다. 표준화에서 표상하는 문법은 기본적으로 개별언어로부터 독립적인 언어 보편성을 지향한다.[4)] 하지만 이를 이용하여 특정한 개별언어의 실제 텍스트에 주석(annotation)을 부착하기 위해서 가장 먼저 할 일은, 각 개별언어에서 시간을 표현하는 언어단위(linguistic unit)를 찾아내는 것이다. 대개 이러한 언어단위는 1개 이상의 어휘로 구성되는 경우가 많고, 언어 규칙으로 생성되는 부분과 단순히 목록으로 제시되어야 할 부분이 혼재하며, 중의성을 가질 가능성이 크다. 따라서 시간을 표현하는 언어단위를 추출하고 그것을 정규화된 방식으로 표상하려면, 이러한 단위의 특성을 규명하는 것이 급선무다.

프랑스어의 시간개체 표현하는 언어단위의 검색과 정규화에 대한 선행 연구로는 Gross (2002)와 Bittar (2008)가 있다. 전자는 시간개체 표현 중 시지점(temporal point)를 나타내는 시각(이하, TIME)과 날짜(이하, DATE) 표현에 해당하는 언어단위를 검색할 수 있도록 [그림1]과 같은 오토마타를 제시했다.[5)]

이 연구가 TIME과 DATE 표현의 대부분에 대해 전체적인 시각에서 언어단위의 구조를 매우 상세히 제시하였다는 장점이 있으나, 지속기간(이하, DURATION)과 반복(이하, SET) 표현을 처음부터 연구대상에서 포함하지 않았고, 각 유형에 따라 정규화된 속성값을 부여하기에는 어려움이 있다. 반면 Bittar (2008)은 Gross (2002)의 연구 결과를 바탕으로 DURATION과 SET 표현을 추가하고, 4개 유형을 <TIMEX3>로 표상하고자 했다.[6)] 하지만

3) 2007년부터 작성되기 시작한 ISO-TimeML은 2008년 9월 제안서 수정이 완료되어 DIS 단계(각 회원국이 가부를 결정하는 투표 단계)에 있다 (ISO-TimeML 2008). 2009년 투표가 완료되면 국제 표준으로 선포된다.

4) 의미주석이 언어 보편성을 지향하지만, 이를 만드는 사람들이 알고 있는 언어에 대한 지식을 넘어서기는 쉽지 않다. 영어사용자 중심으로 만들어진 ISO-TimeML은 이러한 문제점을 극복하기 위해 이태리어, 도이치어, 한국어, 중국어 등 개별언어에 적용하여 검증하는 단계를 거치고 있다 (ISO-TimeML 2008).

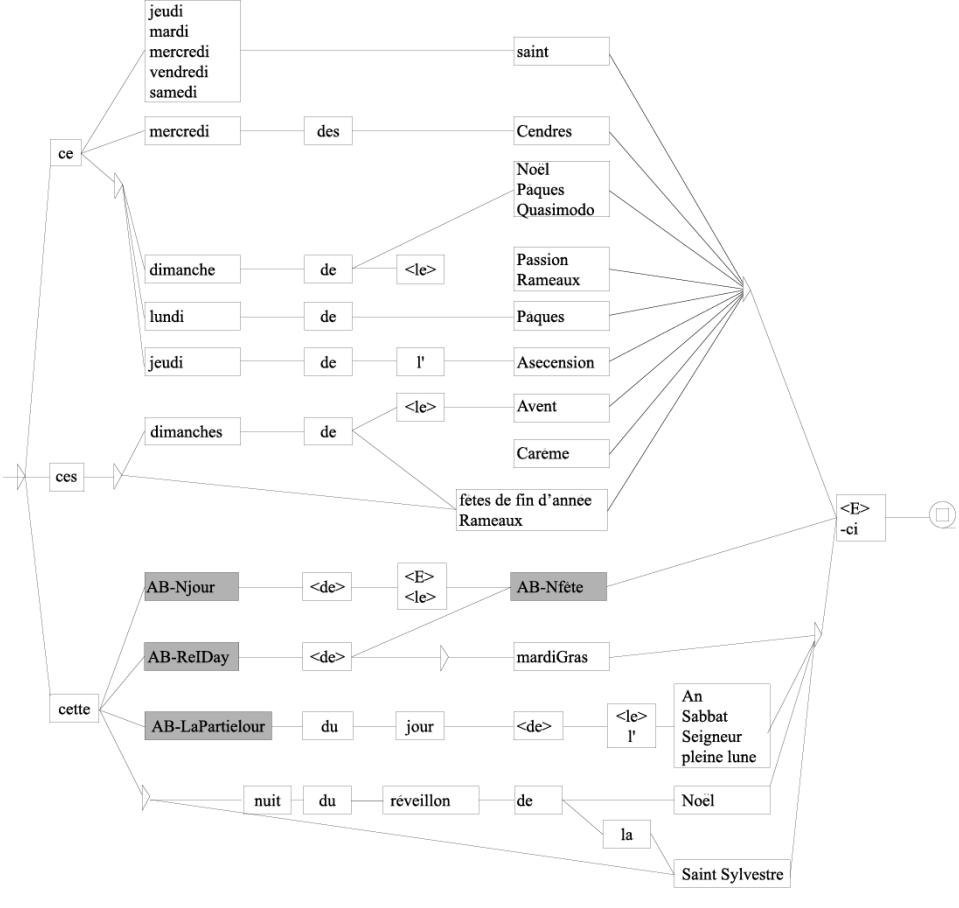
5) 시간개체 표현 유형인 TIME, DATE, DURATION, SET은 <TIMEX>에서 분류하는 메타용어이며 그 정의와 특성은 이 논문 2.1절을 참조하라.. [그림1]은 Unitex로 볼 수 있는 Gross (2002)의 축제일 표현 오토마타 중 하나다.

6) Bittar (2008)은 Gross (2002)의 결과를 포함하여 90개의 오토마타를 제시하였다.

Bittar (2008) 연구의 주요 주석 대상은 신문기사이므로, DURATION과 SET의 언어단위의 구조를 철저하게 제시하지는 못해, 일반 텍스트에 적용하기에는 구성요소의 구조와 목록에 결여된 부분이 많다.

본 연구의 목적은 선행 연구의 한계를 보완하여 프랑스어의 시간개체 표현을 자동으로 검색할 수 있는 오토마타를 설계할 수 있도록 ① 시간을 나타내는 언어단위의 구조를 밝히고, ② 구조를 생성하는 규칙 및 구성요소의 목록을 구축하는 데 있다. 이를 바탕으로 향후 연구에서 자동 검색과 정규화가 이루어질 수 있는 기반을 마련하고자 한다.

이 논문의 구성은 다음과 같다. 2절에서는 시간개체 표현을 표상하는 <TIMEX3>의 특성을 살펴보고, 3.1절에서는 유형별로 프랑스어 시간개체 표현의 대표적인 예를 <TIMEX3>로 정규화할 가능성을 검토해 본다. 3.2절에서는 프랑스어의 시간개체 표현을 자동으로 추출하



[그림 1] 축제일 표현 언어단위를 표상한 오토마타

고, 정규화된 값으로 기술하기 위한 기초 연구로, 시간을 표현하는 언어단위의 구조와 구성요소를 분석해 문자열 틀(template)을 구축할 수 있도록 한다. 마지막 4절에서는 이 논문의 연구 내용을 바탕으로 이루어질 향후 연구 방향을 살펴보겠다.[7)]

2. 시간개체 표현의 표상

2절에서는 ISO-TimeML (2008)의 시간개체 표현 표상을 위한 주석 방식인 <TIMEX3>와 <TLINK>를 소개한다.[8)] 2009년 버전이 2007년-2008년에 나온 이전 버전에 비해 가장 크게 달라진 점은, 이 논문의 (3)처럼 주석된 텍스트와 주석을 물리적으로 분리함으로써, 시간성 표현 이외에도 다양한 층위에서 주석을 할 수 있도록 한 점이다. 이 논문에서는 설명의 편의를 위해 ISO-TimeML (2008)에 소개된 속성의 정의에 (R1)-(R23)과 같은 번호를 붙이고, 값의 순서도 조정했다. <TIMEX3>의 정의에 필요한 속성값의 상세한 예는 ISO-TimeML에 포함되어 있지 않으므로 TIDES (2005)를 참조하였다.

2.1 〈TIMEX3〉

<TIMEX>는 시간선 상에 표시할 수 있는 시간개체 표현의 시지점과 시구간(temporal interval)을 나타내는 정규화 방식을 규정한다. 따라서 시간을 나타내는 표현임에는 분명하나, 시간선 상에 명시적으로 표시할 수 없는 '*longtime ago, one day, shortly*'와 같은 부정적인(indefinite) 표현은 주석의 대상에서 제외한다.[9)]

attributes ::=	id anchor [pred] type (value | valueFromFunction) [functionInDocument] [beginPoint] [endPoint] [quant] [freq] [temporalFunction] [mod] [anchorTimeID] [comment]	(R1)
id ::=	ID {tid ::= TimeID TimeID ::= t<integer>}	(R2)

7) 다양한 유형의 시간개체 표현을 추출할 수 있는 적절한 말뭉치는 찾기 힘들다. 따라서 앞선 두 연구에서는, 연구자가 프랑스 모국어 화자였으므로 모국어 직관을 이용하였다. 본 연구에서는 이를 보완하기 위해 참고문헌 1차 자료에 제시한 프랑스어 문법책에 기술된 시간개체 표현 및 예문과 시간개체 표현을 구성하는 어휘를 대상으로 불-불, 불-한 / 한-불, 불-영 / 영-불 사전에서 예를 수집하였다.

8) ISO-TimeML에서 다루는 시간성 표현은 동사, 서술성 명사 및 형용사 등 '사건(<EVENT>의 주석 대상) 표현'과 '시간개체 표현 간 또는 시간개체 표현과 사건 등의 관계(<LINK>의 주석 대상)'도 포함한다. 사건(<EVENT>)의 속성값이 프랑스어에 적용 가능한지 살펴본 논문은 윤&손 (2008)을 참조하라.

9) 'one day'는 '하루'라는 시간의 양을 나타내는 표현이 아닌 '어느 날'의 뜻을 갖는 경우다.

anchor ::=	IDREF {IDREF ::=(token<integer>)*}	(R3)
pred ::=	CDATA	(R4)
type ::=	'DATE' | 'TIME' | 'DURATION' | 'SET'	(R5)
value ::=	Duration | Date | Time | WeekDate | WeekTime | Season | PartOfYear | PaPrFu	(R6)

<TIMEX3>는 시간개체 표현의 속성을 (R1)처럼 정의한다. 첫 줄에 표시한 ① 각 시간개체 표현에 할당된 고유번호(id, (R2)), ② 텍스트 문자열에서 시간개체 표현이 점유하는 위치(anchor, (R3)), ③ 시간개체 표현의 기본형 값(pred, (R4))은 시간개체 표현의 물리적 속성을 규정한다. ②와 ③은 (ISO-TimeML 2008)에서 처음 도입된 것인데, anchor는 띄어쓰기를 기준으로 토큰(token)을 구분한 후, 각 토큰에 0-N까지 번호를 매긴 후, 시간개체 표현에 해당되는 토큰 번호를 표시하는 것이다. pred는 수의적인 요소로서 해당 시간개체 표현의 기본형을 표시한다.[10)]

둘째 줄에 나타낸 ④ 시간개체 표현의 유형(type, (R5)), ⑤ 시간개체 표현의 속성값(value, (R6)) 또는 그 속성값을 계산하는 기능(valueFromFunction, (R13))은 필수적 요소다.[11)] 시간개체 표현의 유형은 (R5)처럼 DATE, TIME, DURATION, SET으로 분류하며, 시지점 또는 시구간으로 환원하여 그 값을 표현할 수 있고, 유형에 따라 값의 속성이 상이하다.[12)]

시지점로 나타낼 수 있는 것은 TIME과 DATE이며, 1일을 기준으로 그 이하는 TIME, 그 이상은 DATE로 분류한다. 정규화 표현의 예를 살펴보면, 한국어 '08년 10월 24일 오후 8시 반', 영어 '*at 8:30 p.m., Oct. 24th 2008*'와 프랑스어 '*à huit heures et demie, 24 octobre 2008*'의 정규화된 값은 모두 동일하게 val="2008-10-24T20:30:00"로 표현한다.

beginPoint ::=	IDREF {beginPoint ::= TimeID}	(R7)
endPoint ::=	IDREF {endPoint ::= TimeID}	(R8)
quant ::=	CDATA	(R9)
freq ::=	Duration	(R10)

10) 기본형을 표시하는 pred는 시간개체 표현보다는 굴절이나 활용 변화를 하는 동사가 나타내는 '사건(<EVENT>)'에 더 유용하다. 예를 들어 한국어 '사랑했다'는 pred='사랑하다'로, 영어 'taught'은 pred='TEACH'로 된다.

11) valueFromFunction은 추후 수의적인 요소인 temporalFunction(R14) 및 anchorTimeID(R15)와 함께 설명하겠다.

12) 이전 버전과는 달리 <TIMEX3>에서는 시구간을 시간개체 표현의 기본값으로 설정한다. 시구간의 값이 극소화되면, 시지점으로 나타나게 된다.

DURATION은 ⑥ 시작점(R7)과 ⑦ 종결점(R8)으로 규정할 수 있는 시구간으로 표현하는데, 'P(T)n시간단위' 형식으로 기술한다.[13] 예를 들어, '1년, 2달, 3분' 등의 지속기간은 각각 'P1Y, P2M, PT3M' 등으로 나타낸다.

SET 유형에서 시간개체 표현의 양을 표시하는 장치로는 반복단위 1회의 값(value, (R6)), ⑧ 반복단위의 발생 빈도를 나타내는 양화사(quant, (R9)), ⑨ 반복단위 기간 내 사건이 발생하는 횟수(freq, (R10))를 표시한다. 예를 들어 '매주 2일'은 value="P1W", quant="EVERY" freq="2D"처럼 표현된다.[14]

functionInDocument ::= 'CREATION_TIME' | 'EXPIRATION_TIME' | 'MODIFICATION_TIME' |'PUBLICATION_TIME' | 'RELEASE_TIME' | 'RECEPTION_TIME' | 'NONE' {default, if absent, is 'NONE'} (R11)

⑩ 문서 생명주기 시각(functionInDocument, (R11))은 텍스트의 작성, 소멸, 변경, 인쇄, 배포, 접수 시간 등을 나타내며, 당연값(default)은 'NONE'이다.[15]

13) P는 시간개체 표현 유형의 DATE, PT는 TIME를 표현하며, n은 수이고, 시간단위를 나타내는 약어를 정리하면 다음과 같다 (TIDES 2005).

약어	시간단위	약어	시간단위	약어	시간단위
MA	Million years	H1	1st Half Year	MO	MOrning
KA	Thousand years	H2	2nd Half Year	MI	MId-day
CE	Century	Q1	1st Quarter	AF	AFternoon
DE	Decade	Q2	2nd Quarter	EV	EVening
Y	Year	Q3	3rd Quarter	NI	NIght
FY	Fiscal Year	Q4	4th Quarter	PM	PM
SP	SPring	M	Month	DT	DayTime
SU	SUmmer	W	Week	H	Hour
FA	FAll	WE	WeekEnd	M	Minute
WI	WInter	D	Day	S	Secnond

14) <TIMEX2>에 비해 <TIMEX3>에서 정교화된 시간개체 표현 유형이 '반복'이다.

15) 앞서 소개한 바와 같이 시간개체 표현을 정규화하는 연구는 일정 자동관리를 위해 시작했다. 이후 MITRE group 등에서 시간개체 표현 및 사건(<EVENT>)의 자동추출 및 정규화로 확장하면서 활용성이 높은 신문 기사와 같이 특화된 텍스트를 대상으로 삼았다. 이에, functionInDocument는 주로 신문기사의 생명주기(life cycle)와 관련된 시각을 표현한다. 따라서 전통적으로 언어학에서 관심을 갖는 직접인용이나 간접인용 등에서 중요한 기능을 하는 발화시(speech time)나 참조시(reference time)와는 차이가 있다.

mod ::= ‘BEFORE’ | ‘AFTER’ | ‘ON_OR_BEFORE’ | ‘ON_OR_AFTER’ (R12)
|‘EQUAL_OR_MORE’ | ‘EQUAL_OR_LESS’ | ‘MORE_THAN’ |
‘LESS_THAN’
|‘START’ | ‘MID’ | ‘END’ | ‘APPROX’

ⓙ 수식어(mod(R12))[16]는 시간선 상에 적확한 값으로 표현할 수 없는 시간개체 표현의 기간, 양, 양태 등에 대한 추가적인 정보를 기술한다. mod의 하위부류로는 ⓐ 시간개체 표현이 특정한 시지점의 전후관계를 표시하는 전(BEFORE), 후(AFTER), 이전(ON_OR_BEFORE), 이후(ON_OR_AFTER), ⓑ 지속기간의 양을 나타내는 이상(EQUAL_OR_MORE), 이하(EQUAL_OR_LESS), 초과(MORE_THAN), 미만(LESS_THAN), ⓒ 지점과 지속기간을 모두 수식하는 것으로는 시간개체 표현의 국면을 표시하는 시작(START), 중간(MID), 끝(END)과 근사값(APPROX)이 있다. 피수식 시간개체 표현의 특성에 따른 각 값과 그 구체적인 예는 [표 2]와 같다.[17]

[표 2] 시간개체 표현 mod의 값

특성	값	예
points	BEFORE	The site opened *more than* 5 days ago.
	AFTER	The site opened *less than* 5 days ago
	ON_OR_BEFORE	The site opened *no less than* 5 days ago
	ON_OR_AFTER	The site opened *no more than* 5 days ago
durations	LESS_THAN	The site will be open *nearly* 5 days long
	MORE_THAN	The site will be open *more than* 5 days.
	EQUAL_OR_LESS	The site will be open *no more than* 5 days.
	EQUAL_OR_MORE	The site will be open *at least* 5 days.
points, durations	START	The shop opened in the *early* 1960s.
	MID	The site will be open in *mid*-February.
	END	The shop closed at the *end* of 1980s
	APPROX	The site will be open *around* 5:00 pm. The site will be open *about* 5 days.

16) <TIMEX2>부터 설정한 mod는 Modifier의 준말이다.

17) [표 2]는 TIDES (2005:37-38)의 [Table 4-9]에서 예문을 수정 보완한 것이다. 하지만 APPROX은 시지점의 근접성을 나타내는 것과, 시구간의 대략적인 양을 나타내는 것을 구분할 필요가 있다. 이 논문에서는 이를 각각 APPROX_POINT, APPROX_INTERVAL로 지칭하겠다.

temporalFunction ::=	'true' | 'false' {default, if absent, is 'false'} {temporalFunction ::= boolean}	(R13)
anchorTimeID ::=	IDREF {anchorTimeID ::= TimeID}	(R14)
valueFromFunction ::=	IDREF {valueFromFunction ::= TemporalFunctionID TemporalFunctionID ::= tf<integer>}	(R15)

2개 이상의 시간개체 표현 간의 시간순서와 관련해서 수의적 요소인 ⑪ 시간계산 기능의 작동 여부(temporalFunction, (R13))와 ⑫ 시간계산이 작동하는 기준시점(anchorTimeID, (R14))을 통해 표시한다. temporalFunction이 '참'이면, 특정한 기준시점을 설정하고, 기준시점으로부터 시간개체 표현 값을 계산한 기능(valueFromFunction, (R15))이 작동한다.[18] 구체적으로 영어의 예를 살펴보자.[19]

(3) John taught from September to December last year . (written on Oct. 24 2008)

토큰번호 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

<SIGNAL id="s1" anchor="token2" pred="FROM"/>
<TIMEX3 id="t1" anchor="token3" pred="SEPTEMBER" type="DATE" value="xxxx-09"/>
<SIGNAL id="s2" anchor="token4" pred="TO"/>
<TIMEX3 id="t2" anchor="token5" pred="DECEMBER" type="DATE" value="xxxx-12"/>
<TIMEX3 id="t3" anchor="token6 token7" pred="LAST_YEAR" type=DATE "valueFromFunction="2007" temporalFunction="true" anchorTimeID="t4"/>
<SIGNAL id="s3" anchor="token11" pred="ON"/>
<TIMEX3 id="t4" anchor="token12 token 14" pred="OCT._24_2008" type="DATE" value="2008-10-24" functionInDocument="CREATION_TIME"/>
<TIMEX3 id="t5" anchor="type="DURATION" value="P4M" beginPoint="t1" endPoint="t2" temporalFunction="true"/>

18) 이외에도 주석자가 자신의 의견을 기록할 수 있는 'comment::= CDATA'라는 정의가 있으나, 이 논문의 초점을 모으기 위해 <TIMEX>, <SIGNAL>, <TLINK> 소개에서 모두 제외한다.

19) (3)은 ISO-TimeML (2008:22)의 예를 변형한 것이다. <SIGNAL>은 시간개체 표현과 함께 텍스트에 나타나는 'on, in, to before, when, while, -, /' 등 시간 전치사, 접속사, 기호를 분리하여 다음과 같이 표시한다 (TimeML 2006:32).

attributes ::=	id anchor[pred] [comment]	(R16)
id ::=	ID {sid ::= SignalID SignalID ::= s<integer>}	(R17)
anchor ::=	IDREF {IDREF ::=(token<integer>)*}	(R18)
pred ::=	CDATA	(R19)

(e3)에서 시간기호는 '*from*(=s1)', '*to*(=s2)', '*on*(=s3)'이며, 시간개체 표현은 토큰으로 나타난 '*September*(=t1)', '*December*(=t2)', '*last year*(=t3)', '*Oct. 24 2008*(=t4)'와, 구체적인 토큰으로 표시되지 않는 't5(='t1'을 시작점으로 하고 't2'를 종결점으로 하는 '4개월(=P4M)'이라는 기간)'이다. 시지시소(time deixis) '*last year*(=t3)'은 문서작성 날짜인 't4(=2008-10-24)'를 기준시점으로 하여 1년을 빼는 시간계산(-1Y)을 한 값(=2007)을 갖는다.

2.2 〈TLINK〉

<TLINK>는 시간개체 표현 간 맺고 있는 시간순서를 표상한다.[20]

attributes ::=	relType (eventID | timeID)(relatedToEvent | relatedToTime) [signalID] [id] [origin] [comment] [syntax]	(R20)
relType ::=	'BEFORE' | 'AFTER' | 'INCLUDES' | 'IS_INCLUDED' | 'DURING' | 'SIMULTANEOUS' | 'IAFTER' | 'IBEFORE' | 'IDENTITY' | 'BEGINS' | 'ENDS' | 'BEGUN_BY' | 'ENDED BY' | 'DURING INV'	(R21)
timeID ::=	IDREF {timeID ::= TimeID}	(R22)
relatedToTime ::=	IDREF {relatedToTime ::= TimeID}	(R23)

<TLINK>의 가장 중심되는 속성은 필수적 요소인 ① 사건이나 시간개체 표현 간 맺는 관계 유형(relType)이며, (R21)처럼 14개로 분류한다.[21] ② 이러한 관계를 맺기 위해서는 2개의 사건이나 시간개체 표현이 필수적이다. 즉 초점이 되는 사건이나 시간개체 표현(eventID | timeID)와 이와 관련된 사건이나 시간개체 표현(relatedToEvent | relatedToTime)이다. 수의적인 요소로는 사건이나 시간개체 표현과 관련된 시간기호(signalID)가 있다면 표시된다. <TLINK>의 고유번호는 수의적인 요소다. (3)에서 살펴본 시간개체 표현 간의 시간순서를 살펴보자.

---

20) <TIMEX2>에서는 시간개체 표현 간의 순서를 'anchor_dir'이라는 속성으로 표현했으나, TimeML에서는 이를 <TIMEX3>와 <TLINK>로 분리하여 다양한 시간성 표현 간의 순서와 방향성을 표현한다. <TLINK>는 <TIMEX>와 <EVENT>와의 관계를 표상할 수 있으나 이 논문의 논의에서 벗어나는 부분이므로 정의문 소개에서 이 부분((R16)에 나타난 'eventID', 'relatedToEvent')의 설명을 제외하겠다.

21) 관계 유형 14개는 기본적으로 Allen, J. 1984. "A General Model of Action and Time", *Artificial Intelligence* 23, 2. (ISO-TimeML, 2009:31-32 재인용)의 13개 분류를 따르며, B라는 기간 내에 A가 반복되는 유형인 DURING_INV를 추가한 것이다. Allen과 ISO-TimeML의 메타용어 비교는 윤&손(2008)을 참조하라.

(3')
<TLINK timeID="t1" signalID="s1" relatedToTime="t5" relType="BEGINS"/>
<TLINK timeID="t2" signalID="s2" relatedToTime="t5" relType="ENDS"/>
<TLINK timeID="t5" relatedToTime="t3" relType="IS_INCLUDED"/>
<TLINK timeID="t3" relatedToTime="t4" relType="BEFORE"/>
<TLINK timeID="t5" relatedToTime="t4" relType="BEFORE"/>

't5(=*from September to December*)'를 중심으로 했을 때, 't1(=*September*)'은 시작(=BEGINS)을, 't2(=*December*)'는 끝(=ENDS)을 나타낸다. 이때 's1(=*from*)'과 's2(=*to*)'는 각각 't1'과 't2'에 연결되어 있다. 또한 't5'는 't3(=*last year*)'에 포함되고, 't4(=2008-10-24)'보다 선행한다.[22)]

3. 프랑스어 시간개체 표현의 언어단위 분석

프랑스어 텍스트에서 시간개체 표현을 자동으로 검색 및 추출하고, 이 논문 2절에서 기술한 정규화된 값을 부여하기 위해서는, 우선 프랑스어에서 시간을 표현하는 언어단위의 구성요소를 유형별로 분석해야 한다. 이러한 시간개체 표현을 자동으로 추출하는 방법은 ① 형태분석(morphological analysis)을 거쳐 기본형, 굴절, 품사 정보가 태깅된 텍스트에서 찾는 경우와 ② 이러한 정보가 없는 원자료(raw text)에서 바로 찾는 경우에 따라 차이가 있다. 전자는 검색에 필요한 사전의 크기가 작고 1차로 검색된 결과의 중의성이 낮은 반면, 형태소 분석기를 탑재하고 분석 단계를 거쳐야 하므로 검색 시간이 오래 걸리고 메모리를 많이 소요한다. 반면에 후자는 분석 단계 없이 텍스트 표면형에 문자열 일치(string match) 기법을 사용하므로 검색 속도가 빠르나, 사전이 크고 검색 결과의 중의성이 높다. 따라서 목적, 검색 대상의 특성과 탑재할 기기의 성능, 선행 연구기반에 따라 적합한 방식을 채택하게 된다. 본 연구에서는 형태소 분석기를 사용하지 않는 후자에 필요한 프랑스어 시간개체 표현의 언어적 특성을 중점적으로 살펴보겠다.

3.1절에서는 4개 유형에 따라 프랑스어 시간개체 표현을 <TIMEX3>로 기술하여 메타언어로서의 표현력을 검증해 보고, 3.2절에서는 문자열 일치에서 검색의 대상이 되는 시간개체 표현의 틀을 구성하는 언어단위의 특성을 분석한다.

22) 논리적으로 유도가 가능한 정보는 중복적이므로 텍스트에 주석으로 표시하지 않아도, 시간순서 해석규칙을 통해 획득할 수 있다.

3.1 〈TIMEX3〉를 이용한 프랑스어 시간개체 표현의 정규화

TIME, DATE, DURATION, SET에 해당하는 프랑스어 시간개체 표현을 <TIMEX3>로 정규화해 보자.[23] (4)-(7)에 표현된 텍스트를 작성하는 현재 날짜와 시각이 서양력 2008년 10월 24일 오후 6시이며,[24] 작성자와 각 문장의 행동주가 같은 시간대를 사용하는 지역에 있다고 가정하자.

3.1.1 TIME, DATE

(4) a. Il est arrivé à 15h 30.
b. Mis à jour le 01.10.08 : 22h23
c. Il est arrivé il y a une heure.
d. Il est arrivé hier au soir.

a'. <SIGNAL id="s1" pred="À"/>
<TIMEX3 id="t1" pred="15H_30" type="TIME" value="15:30"/>

b'. <TIMEX3 id="t1" pred="LE_01.10.08_: _22h23" type="TIME" value="2008-10-01T22:23"/>

c'. <SIGNAL id="s1" pred="IL_Y_A"/>
<TIMEX3 id="t1" pred="UNE_HEURE" type="TIME" valueFromFunction="2008-10-24T17:00" temporalFunction="true" anchorTimeID="t2"/>
<TIMEX3 id="t2" pred="" type="TIME" value="2008-10-24T18:00" functionInDocument="CREATION_TIME"/>

d'. <TIMEX3 id="t1" pred="HIER_AU_SOIR" type="TIME" valueFromFunction="2008-10-23TEV"temporalFunction="true" anchorTimeID="t2"/>
<TIMEX3 id="t2" pred="" type="TIME" value="2008-10-24T18:00" functionInDocument="CREATION_TIME"/>

(4) 중 a와 b는 TIME 또는 TIME+DATE의 절대값을 표현한 것이고, c와 d는 '*il y a une heure*'와 '*hier*' 같은 시지시소를 포함하고 있으므로. 현재 날짜와 시각을 기준시점으로 '1시간'과 '1일'을 빼는 시간계산 기능이 필요하다.

---

23) 지면 관계상 <TIMEX3>로 정규화된 표현에서 문장 내 토큰의 순서를 나타내는 anchor="token#"는 생략한다.
24) 텍스트 작성 시각은 (4')-(7')에 모두 표현할 수 있으나 지면상 줄이고 temporalFunction="true"인 경우만 표시한다. 작성자와 행동주가 다른 시간대를 사용하는 경우라면 GMT를 기준으로 시간을 계산해야 한다. <TIMEX3>는 시간대 계산을 표현할 수 있는 기능이 있다.

(5) a. Il écrivît dans les années 30.
b. La famille se réunit pour le Jour de l'an.
c. Il est né il y a à peu près soixante-dix ans.
d. Il va travailler au FNAC à partir du lundi prochain.

a'. <SIGNAL id="s1" pred="DANS"/>
<TIMEX3 id="t1" pred="LES ANNEES 30" type="DATE" value="193"/>
b'. <TIMEX3 id="t1" pred="LE JOUR DE L'AN" type="DATE" value="XXXX-01-01"/>
c'. <SIGNAL id="s1" pred="IL Y A"/>
<TIMEX3 id="t1" pred="A PEU PRES SOIXANTE-DIX ANS" type="DATE" mod="APPROX_INTERVAL" valueFromFunction="1938" temporalFunction="true" anchorTimeID="t2"/>
<TIMEX3 id="t2" pred=" type="DATE" value="2008-10-24" functionInDocument="CREATION_TIME"/>
d'. <SIGNAL id="s1" pred="À PARTIR DU'"/>
<TIMEX3 id="t1" pred="LUNDI PROCHAIN" type="DATE" mod="ON_OR_AFTER" valueFromFunction="2008-10-27" temporalFunction="true" anchorTimeID="t2"/>
<TIMEX3 id="t2" pred=" type="DATE" value="2008-10-24" functionInDocument="CREATION_TIME"/>

절댓값으로 표시할 수 있는 4(a-b)의 *'le 01.10.08'*이나 동일한 값을 갖는 다른 표현인 *'le 1er octobre 2008'*은 모두 value='2008-10-01'이 된다. (5a)에서 1930년대를 지칭하는 *'les années 30'*은 통상 4자리로 표시되는 연도에 3자리 값을 가진 value="193"으로 표현한다.[25] (5b)에서는 *'Jour de l'an'*과 같이 날짜 절댓값을 갖는 축제, 경축일, 국경일 등 고유명사가 value="XXXX-01-01"로 표현된다. 이때 연도를 나타내는 4자리가 XXXX로 표시된 것은 특정한 해가 아님을 표시하며, *'le Jour de l'an cette année'*는 문서작성 시간에서 계산한 값을 가져와 value="2008-01-01"이 된다. 마찬가지로 시지시소를 포함한 (5c)와 (5d)는 문서작성 날짜를 기준시점으로 계산을 통해 값을 갖는다. 하지만 (5c)에서 그 값이 근삿값임을 나타내는 *'à peu près'*는 mod="APPROX"으로 표시하고, (5d)에서 출발점임을 알리는 *'à partir du'*는 mod="ON_OR_AFTER"로 표현한다.

25) <TIMEX3>에서 구체적인 예가 없어 <TIMEX2>를 참조하였다 (TIDES 2005:14-15). 이와 같은 원칙에서 세기는 2자리로 표현하는데, 예를 들면 *'17e siècle'*는 value="16"가 된다.

3.1.2 DURATION

(6) a. Il y est resté là pour au moins 3 heures.
b. Il est en vacances depuis quinze jours.
c. Il travaille ici depuis ce lundi.
d. La Renaissance se développe du XVe siècle à la fin du XVIe siècle.

a'. <SIGNAL id="s1" pred="POUR"/>
<TIMEX3 id="t1" pred="AU MOINS 3 HEURES" type="DURATION" value="PT3H" mod="EQUAL_OR_MORE"/>

b'. <SIGNAL id="s1" pred="DEPUIS"/>
<TIMEX3 id="t1" pred="QUINZE JOURS" type="DURATION" value="P15D" beginPoint="t2" endPoint="t3" temporalFunction="true"/>
<TIMEX3 id="t2" pred="" type="DATE" valueFromFunction="2008-10-10" temporalFunction="true" anchorTimeID="t3"/>
<TIMEX3 id="t3" pred="" type="DATE" value="2008-10-24" functionInDocument="CREATION_TIME"/>

c'. <SIGNAL id="s1" pred="DEPUIS"/>
<TIMEX3 id="t1" pred="" type="DURATION" value="P5D" beginPoint="t2" endPoint="t3" temporalFunction="true"/>
<TIMEX3 id="t2" pred="CE LUNDI" type="DATE" valueFromFunction="2008-10-20" temporalFunction="true" anchorTimeID="t3"/>
<TIMEX3 id="t3" pred="" type="DATE" value="2008-10-24" functionInDocument="CREATION_TIME"/>

d'. <SIGNAL id="s1" pred="DU"/>
<TIMEX3 id="t1" pred="" type="DURATION" value="P2CE" beginPoint="t2" endPoint="t3" temporalFunction="true"/>
<TIMEX3 id="t2" pred="XVe siècle" type="DATE" value="14"/>
<SIGNAL id="s1" pred="À la"/>
<TIMEX3 id="t3" pred="FIN DU XVIe siècle" type="DATE" value="15" mod="END"/>

(6)에서 단순한 기간의 표시는 (6a')에서처럼 3시간은 value="PT3H"으로, 지속기간이 3시간 이상임을 나타내는 '*au moins*'은 mod="EQUAL_OR_MORE"로 표현한다. 시작점과 종결점을 알 수 있는 경우, beginPoint와 endPoint를 이용하여 지속기간을 표현한다. (6b')에서 지속기간은 '*quinze jours*'이며 종결점은 문서작성 날짜이고, 시작점은 이보다 15일 전인 '2008년 10월 10일'이라는 것은 시간계산에 의해 추출된다. 지속기간이 텍스트에 표면형을 갖지 않는 경우(pred="")도 시작점과 종결점을 이용하여 표현할 수 있다. (6c')에서는 종결점(=t3)이 문서작성 날짜이고, 그 주 월요일이라는 시지지소 '*ce lundi*'를 시작점(=

t2)으로 계산하여 지속기간의 value=“P5D”인 t1으로 표현한다. (6d')에서 시작점은 ‘*XVe siècle*(=t2)’이고 종결점은 ‘*fin du XVIe siècle*(=t3)’이므로 각각을 DATE 유형으로 표시한 후, 지속기간의 value=“P2CE”인 t1으로 표현한다.

3.1.3. SET

(7) a. Il se rend visite à sa grand-mère le lundi.
b. Il publie un magazine hebdo.
c. Il se rend visite à sa grand-mère 2 fois par mois.
d. Ce magazine se publie bi-mensuellement.
e. Les conférences ont lieu le 1er jeudi de chaque mois.

a'. <TIMEX3 id=“t1” pred=“LE LUNDI” type=“SET” value=“XXXX-WXX-1” quant=“EVERY”/>
b'. <TIMEX3 id=“t1” pred=HEBDO” type=“SET” value=“P1W” quant=“EVERY”/>
c'. <TIMEX3 id=“t1” pred=“2 FOIS PAR MOIS” type=“SET” value=“P1M” quant=“EVERY” freq=“2X”/>
d'. <TIMEX3 id=“t1” pred=“BI-MENSUELLEMENT” type=“SET” value=“P1M” quant=“EVERY” freq=“2X”/>
e'. <TIMEX3 id=“t1” pred=“LE 1ER JEUDI DE CHAQUE MOIS” type=“SET” value=“XXXX-XX-W1-4” quant=“EVERY”/>

type=“SET”을 나타내는 (7a')에서 ‘*le lundi*’는 한 주의 첫 번째 날을 뜻하는 value=“XXXX-WXX-1”와 그것이 주마다 이루어짐을 나타내는 quant=“EVERY”로 표현한다. 특정한 요일을 나타내지 않아도 좋은 (7b')에서는 value를 기간의 단위인 “P1W”으로 표현한다. (7c)와 (7d)에서 ‘*2 fois par mois*’와 ‘*bi-mensuellement*’는 텍스트 표면형은 다르나 ‘매달 2회’라는 같은 값(value=“P1M” quant=“EVERY” freq=“2X”)으로 표현한다. (7e)는 SET의 가장 복잡한 예다. ‘*le 1er jeudi de chaque mois*’에서 ‘한 달의 첫 번째 목요일’은 value=“XXXX-XX-W1-4”로,[26] 매달 반복됨은 type=“SET” quant=“EVERY”로 나타낼 수 있다.

3.2 시간개체 표현의 언어단위 구조 분석 및 틀 정의

텍스트에서 주석대상이 될 시간개체 표현을 검색하려면, (4)-(7)의 밑줄친 부분과 같은 시간개체 표현의 문자열 틀을 정의하고, 각 틀을 구성하는 성분에 삽입될 수 있는 어휘 사전

26) ‘value=“XXXX-XX-W1-4”’은 이 논문의 저자가 제안한 표현값이다.

이 필요하다. 이때 틀 정의는 오토마타를 정의하여 문자열의 최장일치 검색이 가능하도록 한다. 따라서 틀에는 시간개체 표현인 <TIMEX>의 주석대상뿐만 아니라 이와 밀접하게 관련된 <SIGNAL>도 포함해야 한다.[27)]

3.2.1 TIME과 DATE

앞에서 살펴본 (4)-(5)처럼 나타나는 시각과 날짜 표현의 틀의 구성은 크게 (T1)과 같이 표현할 수 있다.

PointTemplate ::= [PointPreSignal] [PreModifier] (TriggerWithoutTemporalFunction|TriggerWithTemporalFunction) [PostModifier] [PointPostSignal] (T1)

① 필수적인 구성요소로 시지점의 type과 value를 결정하는 중심어휘 표지(lexical trigger)의 속성값을 설정하는 데, '시간계산 기능이 작동되어야 하는 경우(이하, TriggerWithTemporalFunction)'와 그렇지 않은 경우(이하, TriggerWithoutTemporalFunction)로 분류한다. 전자는 value를 결정할 때 anchorTime과 시간계산 기능을 표시해야 한다. 수의적 구성요소로는 ② 시지점을 표현하는 기호인 Signal과, ③ 중심어의 앞과 뒤에서 이를 수식하는 Modifier이다. 이때 문자열 틀에서는 출현 순서에 민감하므로 Signal과 Modifier는 중심어휘 표지를 기준으로 앞과 뒤에 놓일 수 있는 2종류(Pre, Post)로 구분하고, Modifier는 중심어휘 표지의 본래적 의미 특성에 따라 2종류(Point, Interval)로 구분한다.[28)]

3.2.1.1 TriggerWithoutTemporalFunction 언어단위의 기본 구조

시지점을 표현하는 시간개체 표현이 고유값을 갖는 TriggerWithoutTemporalFunction의 경우, 기본적인 틀은 이 논문 4.1절의 (4a-b)처럼 시각과 날짜를 표현하는 특수한 패턴, (5b)처럼 축제나 국경일 등을 지칭하는 고유명사, (5a)처럼 일반 시간명사구로 나타난다. 틀을 만들기 위해, 중심어의 종류를 기준으로 결합할 수 있는 <SIGNAL>의 종류를 구분하면 [표 3]과 같다.[29)]

---

27) 문자열 틀을 규정하는 단계에서부터, 향후 연구에서 검색된 문자열을 자동으로 유형 분류하고, 값을 설정할 수 있어야 함을 염두에 두어야 한다.

28) 2절에서 (R1)과 [표 2]에서 사용한 mod는 시간개체 표현이 갖는 속성값을 나타내고, Modifier는 그것을 나타내는 언어기호를 기칭한다.

29) [표 3]에서 중심어가 고유명사(구)나 일반명사(구)인 시지점 표현은 PointPreSignal을 달리 쓰거나 문맥에 따라서 지속기간을 나타낼 수 있다. 또한 *'fête des Mères'*는 '5월의 마지막 일요일'로 매년 날짜가 바뀌나, 이것은 텍스트 내부에서 기준시점을 갖는 것이 아니라 외부적으로 결정되는 것인 만큼 텍스트 내부에서는 절댓값을 갖는다.

[표 3] TriggerWithoutTemporalFunction의 예 (발췌)

중심어휘표지구분	PointPreSignal	중심어휘 표지	중심어휘 표지분류
특수 패턴	[à]	17:45	TimeFormular DateFormular
	[à]	5:45 pm	
	à	six heures moins le quart du matin	
	[à]	le 24.10.2008	
	[à]	le 24/10/08	
	[à]	vendredi, 24 octobre 2008	
	[à]	le 01.10.08 : 22h23	
고유명사(구)	[à]	la Noël	ProperNoun
	[à]	la fête_des_Mères	
	[à]	le semestre de janvier	
	[à]	la fête nationale	
	[à]	la Seconde Guerre Mondiale	
일반명사(구)		lundi	WeekDate
	dans	la matinée	PartOfDay
	au	mois de mai | printemps	Month | Season
	à	la 15e semaine de l'année	Week
	en	mai | automne | 1989	Month | Season | Year
	dans	les années 30	Decade
	au	20e siècle	Century

이때 TimeFormular, DateFormular, ProperNoun은 패턴 및 어휘 목록을 만들고 시간개체 표현 고윳값을 정의해 주어야 한다. 이 뿐만 아니라 <TIMEX>에서 값을 정의한 WeekDate, PartOfDay, Month, Season, Year 등에 시지점 표현[30]에 해당하는 프랑스어 어휘목록이 필요하며, 이 어휘가 사용되는 문자열 틀을 규정해야 한다.

3.2.1.2 TriggerWithTemporalFunction 언어단위의 기본 구조

시지점을 표현하는 시간개체 표현이 시간계산 기능을 작동하여 그 값을 얻는 TriggerWithTemporalFunction의 경우, 기본적인 틀은 일반 시간명사 / 부사(구)로만 이루어진다. 이러한 시간개체 표현은 (4c-d)와 (5c-d)처럼 시간계산을 할 때 기준시점(AnchorTime)이 문서작성 시각(또는 발화 현재)인 경우와 그렇지 않은 경우로 구분할 수 있

30) [표 3]에서 PointPreSignal과 함께 사용되는 일반명사(구)의 중심어휘 표지 '*matinée, printemps, mai*' 등의 본래적 의미는 '시지점'을 나타낼 수도 '시기간'을 나타낼 수도 있다. 사용되는 문맥에 따라 전자는 TIME또는 DATE 유형으로, 후자는 DURATION으로 분류한다.

다. 또한, 중심어휘 표지 자체의 본래적인 의미 특성에 따라, ‘시지점을 나타내는 표지(이하, PointTrigger)’와 ‘시구간을 나타내는 표지(이하, IntervalTrigger)’로 구분한다.

[표 4] TriggerWithTemporalFunction의 예 (발췌)

기준시점 구분	PointPreSignal	중심어휘 표지	PointPostSignal
기준시점= 문서작성시각		maintenant | présentement	
		aujourd’hui | hier au soir	
	(à)	ce samedi | le mois dernier | l’année prochaine	
	à	l’heure actuel | présent	
	en	ce moment | ce temps même	
두 경우에 모두 속함	il y a	1 heure | 2 jours | 4 semaines	
		1 heure | 2 jours | 4 semaines	avant | après | plus tard
기준시점≠ 문서작성시각		ce matin-là | cette année-là	
	(à)	le lendemain | la veille au soir | la semaine précédente | le jour suivant | le jour d’après	
	à	67 ans | l’âge de 67 ans	
	(	67	)

PointTrigger로, 전자는 ‘*maintenant, hier, après-demain, ce, dernier, prochain, actuel, à présent’*과 같은 시지시소를 포함하고, 후자는 ‘*ce* 시간명사-*là’*나 ‘*lendemain, veille, précédent, suivant, d’après’* 등과 같은 고정적인 표현을 포함하거나, 출생연도를 기준으로 나이를 나타내는 ‘*à 67 ans, à l’âge de 67 ans, (67)’*를 포함한다. 하지만 두 경우 모두에서 ‘*1heure, 2 jours, 4 semaines’* 등처럼 IntervalTrigger가 PointPreSignal ‘*il y a’*나 PointPostSignal ‘*avant, après, plus tard’* 등과 결합해 나타날 수 있으며, 이때 ‘*il y a’*와 ‘*plus tard’*의 기준시점은 텍스트에 달리 명시되지 않으면 문서작성 시간이 된다.

3.2.1.3 MOD

TIME과 DATE의 속성값을 수식하는 Mod는 중심어휘 표지가 PointTrigger인지 IntervalTrigger인지에 따라 구분해야 한다.[31] 전자는 [표 5]와 같이 PointPreSignal을 대체하는 PointPreModifier인 전치사(구)로 표현되며,[32] mod값이 ON_OR_BEFORE 또는

31) TriggerWithoutTemporalFunction인 경우, 중심어휘 표지는 모두 PointTrigger이다.

32) APPROX_POINT인 ‘*à peu près’*는 부사구이므로 다른 전치사(구)인 Modifier와는 다른 특성을 보인다. ‘*à peu près à cette époque* | *à peu près à 2 ans’*처럼 PointPreSignal 앞에서 시지점의 근접성을 나타낸다

ON_OR_AFTER인 경우에만 PointPostModifier로 표현할 수 있다.

[표 5] PointTrigger과 공기하는 Modifier와 mod 값 (발췌)

mod구분	mod 값	PointPreModifier	중심어휘 표지	PointPostModifier
시지점의 전후관계	BEFORE	avant | juste avant | vers	PointTrigger	
	AFTER	après | juste après		
	ON_OR_BEFORE	jusqu'à | jusqu'à la fin de | dès avant | pas plus tard que		au plus tard
	ON_OR_AFTER	depuis | à partir de | dès | dès après | dès le début de		au plus tôt
시지점의 근접성	APPROX_POINT	près de | à peu près | aux environs de | autour de		ou à peu près
시지점/ 시구간의 국면	START	au début de |		
	MID	au milieu de | vers le milieu de		
	END	à la fin de		

IntervalTrigger를 수식하는 mod는 2가지로 구분할 수 있는데, [표 6]에서 볼 수 있는 것처럼 하나는 시구간의 양화 표현이고, 다른 하나는 시지점의 전후관계를 나타내는 것이다. 이때 전자는 독립적으로 사용되는 것이 아니라 '*dans moins de 15 jours, il y a plus ou moins 3 jours, une vingtaine de jours après, d'ici 3 jours à peu près, à peu près 2 ans plus tard, dans un mois au plus tard'*에서처럼 [표 4]와 [표 5]의 PointPreSignal 및 PointPostSignal과 결합하여 사용된다.

[표 6] IntervalTrigger과 공기하는 Modifier와 mod 값 (발췌)

mod구분	mod 값	IntervalPreModifier	중심어휘 표지	IntervalPostModifier
시구간의 양	LESS_THAN	moins de	Interval Trigger	
	MORE_THAN	plus de		
	EQUAL_OR_LESS	pas plus de		au maximum
	EQUAL_OR_MORE	pas moins de		au moins | au minimum
	APPROX_INTERVAL	approximativement | plus ou moins | presque | à peu près | environ | une vingtaine de		environ | à peu près

mod구분	mod 값	IntervalPreModifier	중심어휘 표지	IntervalPostModifier
시지점의 전후관계	BEFORE	avant | dans | dans pas | dans moins de | il y a plus de	Interval Trigger	avant
	AFTER	après | dans | dans plus de | d'ici | d'ici à | il y a moins de		après
	ON_OR_BEFORE	dès avant | il y a au moin de		au plus tard
	ON_OR_AFTER	dès après | il y a pas plus de		au plus tôt

3.2.2 DURATION

앞에서 살펴본 (6)처럼 나타나는 지속기간을 표현하는 틀의 구성은 크게 (T2)과 같이 정의할 수 있다.

DurationTemplate ::= [DurationPreSignal] [IntervalPreModifier] IntervalTrigger [IntervalPostModifier] [DurationPostSignal] | (([StartPointPreSignal | EndPointSignal]) [PointPreModifier] PointTrigger [PointPostModifier] [PointPostSignal])+ (T2)

지속기간은 크게 2가지 방식으로 표현된다. 하나는 (6a-b)처럼 시구간을 이용하여 표현하는 것인데, [표 7]에서와 같이 (6a)처럼 시작점 또는 종결점과 같은 시지점을 알 수 없는 경우와 (6b)처럼 시작점을 추론할 수 있는 경우로 구분할 수 있다. 단 종결점을 추론할 수 있는 '*Il est **dans la 3e année***. | *Il terminera son travail **[en 3 jours** | **dans 3 jours** | **d'ici 3 jours]***' 등의 경우는 [표 8]과 같이 mod 값이 추가된다.

[표 7] IntervalTrigger을 이용한 DURATION 표현 (발췌)

구분	DurationPreSignal	중심어휘 표지	DurationPostSignal
시지점 유추 불가능	pendant | durant | pour | [dans] le courant de | au cours de | dans le cours de	IntervalTrigger	durant
	TOUT	IntervalTriggerSG	
시지점 유추 가능	depuis	IntervalTrigger	

[표 8] IntervalTrigger을 이용한 DURATION와 공기하는 mod (발췌)

구분	mod 값	DurationPreModifier	중심어휘 표지
종결점 유추 가능	BEFORE	en | dans LE Ordinal	IntervalTrigger
	AFTER	dans | d'ici	

DURATION을 표현하는 다른 한 유형은 [표 9]에서 같이 (6c-d)처럼 시작점과 종결점을 표시하는 것이다. 이 유형도 (6c)처럼 시작점과 종결점 어느 하나만 나타낸 경우와 (6d)처럼 두 개 모두 나타낸 경우가 있으며, 이때 시작점과 종결점은 시지점으로 환원될 수 있다. 단, 후자에서 PointPreSignal은 '시작'과 '끝'을 나타낸다. 시작점과 종결점에 대한 명확한 언급이 없으면 당연값은 문서작성 시간(발화현재)이다.[33]

[표 9] PointTrigger을 이용한 DURATION 표현 (발췌)

구분	PointPreSignal	중심어휘 표지
시작점	de | depuis | à partir de | dès | à	PointTrigger
종결점	à | jusqu'à | d'ici [à]	

이상에서 소개한 지속기간 표현을 수식하는 Modifier는 중심어휘 표지의 유형에 따라 [표 5]와 [표 6]에서 살펴본 Modifier와 공기한다.

3.2.3 반복: SET

앞에서 살펴본 (7)처럼 나타나는 반복을 표현하는 틀의 구성은 크게 (T3)과 같이 정의할 수 있다.

```
SetTemplate ::= [PreModifier] ([Freq] [PreModifier] SetUnit | SetAdj&Adv)      (T3)
                [PostModifier]
Freq ::=        Number [IntervalTrigger]
                {Num ≧ 1
                default, if absent, is '1'}
SetUnit ::=     ((par | chaque | LE) (IntervalTriggerSG | PointTiggerSG)) |
                (TOUS LES ([Quant] IntervalTriggerPL | PointTiggerPL))
                {Quant ::= CDATA
                default, if absent, is 'EVERY'}
```

33) *'d'ici (à) samedi, d'ici le 13'*에서처럼 *'d'ici'*는 시작점인 문서작성 시간을 나타내고 *'(à) samedi, le 13'*는 종결점을 표현함으로써 지속기간을 표시한다.

2.1절에서 살펴보았듯이 반복의 의미적 구성요소는 ① 반복단위 1회의 값, ② 반복단위의 발생 빈도(=Quant), ③ 반복단위 기간 내 사건이 발생하는 횟수(=Freq)다. 이것이 반복 틀에서는 [표 10]처럼 크게 2가지 패턴으로 이루어진다. 하나는 (7a,c,e)처럼 ①과 ②를 표현하는 SetUnit과 ③인 Freq으로 구성된 경우와, 다른 하나는 (7b-d)처럼 ①, ②, ③이 어휘 하나에 표현되는 SetAdj&Adv이다. 후자는 '*hebdomadaire, annuel, mensuel, bi-mensuel*' 등 형용사와 이것에서 파생된 부사로 이루어진다. 전자는 '*mois, matin, lundi*'처럼 SetUnit을 구성하는 Trigger가 단수형이고 그 앞에 반복을 나타내는 표지인 '*par, chaque, le*' 등이 있는 것과 Trigger가 복수형이고 그 앞에 집합의 구성 전체를 나타내는 '*tous les, toutes les*' 등과 결합한 것이 있다.

[표 10] SET 표현 패턴 (발췌)

구분		Freq		SetUnit
		Number	IntervalTrigger	
SetUnit	IntervalTriggerSG	15	jours	par mois
		2	fois	chaque matin
	PointTriggerSG	5	heures	le lundi
	IntervalTriggerPL	5	jours	tous les 2 mois
	PointTriggerPL	3	fois	tous les 1ers lundis du mois
SetAdj&Adv		hebdomadaire		
		bi-mensuellement		

Mod로는 [표 6]에서 소개한 IntervalTrigger의 양화 표현 중 APPROX_INTERVAL에서 전치사구 앞에 놓일 수 있는 통사적 특성을 가진 IntervalPreModifier '*approximativement, plus ou moins, presque, à peu près, environ*'와 IntervalPostModifier인 '*à peu près, environ*'만이 공기 가능하다. 특히 IntervalPreModifier는 Freq 앞이나 SetUnit 앞에 위치할 수 있다. 예를 들어 '*à peu près 15 jours par mois, 15 jours par mois à peu près, 15 jours à peu près par mois*'와 같이 표현될 수 있다.

4. 이어가기

본 연구에서는 프랑스어의 시간개체 표현을 자동으로 검색할 수 있는 오토마타를 설계할 수 있도록 ① 시간을 나타내는 언어단위의 구조를 밝히고, ② 구조를 생성하는 규칙 및 구성요소의 목록을 구축하였다. 이를 바탕으로 향후 연구에서 자동 검색과 ISO-TimeML로 정규

화할 수 있는 기반을 마련하고자 하였다. 특히 형태소 분석 단계 없이 텍스트 표면형에 문자열 최장일치 기법을 사용하기 위해 검색의 대상이 되는 시간개체 표현의 틀을 구성하는 언어단위의 특성을 분석했다. 선행연구의 한계를 보완하여, ① 연구대상의 유형을 TIME과 DATE뿐 아니라 DURATION과 SET을 포함하도록 확장하였고, ② 동시에 각 유형의 시간개체 표현을 생성 및 분석할 수 있는 틀 및 그 구성요소 간의 결합관계를 밝혔다. 동시에 후속연구에서 각 틀에 해당하는 시간개체 표현에 속성값을 자동으로 부여할 수 있는 기초 연구를 수행하였다.

이 논문을 기반으로 앞으로 수행해야 할 연구는 많이 남아 있다. 첫째, 본 연구에서 분석한 프랑스어 시간개체 표현의 틀을 실제 다양한 종류의 텍스트에 적용하여 검색의 정확도와 재현율을 평가해야 한다. 둘째, 아무리 정확도가 높더라도 ISO-TimeML과 같은 메타언어를 이용하여 실제 텍스트에서 시간성 표현에 주석을 다는 일을 사람이 수작업으로 할 수는 없다. 따라서 자동으로 검색하고, 유형을 분류하고, 속성값을 할당하는 자동주석 시스템을 만들어야 한다. 이를 위해서 시간을 표현하는 언어단위의 중의성을 해결할 수 있는 기계학습 방법론을 연구해야 하며, 시간계산 기능이 필요한 시간개체 표현의 경우, 텍스트 내부와 발화상황에서 기준시점을 찾아내고, 속성값을 계산할 수 있는 규칙이 필요하다. 셋째, 시간개체 표현을 바탕으로 다문서, 다언어와 같은 분산된 환경에서 나타난 동일한 또는 관련이 있는 사건 간 추론이 가능하도록 일반적인 추론 규칙을 설정해야 한다. 이러한 추론 규칙은 1절에서 살펴봤던 (1)-(2)로부터 [표 1]과 같은 단순한 지식을 추출할 수 있을 뿐 아니라 더욱 복잡한 맥락의 지식을 도출해 낼 수 있을 것이다.

20세기 중후반에 나타난 컴퓨터와 통신망의 등장은 불과 20여 년 만에 '제2의 구텐베르그 혁명'이라고 불릴 만큼 인간 지식의 생성·유통·관리에 완전히 새로운 패러다임을 창출했다. '한 노인이 죽는 것은 도서관 하나가 불타 없어지는 것과 같다.'라고 하는 아프리카 속담은 비단 문자 기록을 통한 지식의 축적이 이루어지지 않는 지역에만 해당하는 것은 아니다. 과유불급(過猶不及)! 옥석을 가릴 수 없는 텍스트가 폭포수처럼 밀려드는 정보화 사회에서는 텍스트 안에 담긴 의미를 추출하고, 이를 통하여 지식을 재구성할 수 있지 않는 한 '한 전문가가 죽는 것은 한 전문 분야가 사라지는 것과 같다.'라는 새로운 버전의 속담이 정보화 사회에도 유효할 것이다. 지식처리의 필요성은 여기에 있고, 텍스트에 담긴 시간 정보를 검색하고 정규화하려는 본 연구는 의미처리를 위한 다양한 노력의 일환이다.

참고문헌

윤애선 , 손현정. 2008. “프랑스어 시간성 표현의 정규화를 위한 메타언어 검증”. 『불어불문학연구』 76, pp.433-467.

Bittar, A. 2008. Annotation des informations temporelles dans des textes en fraçais. in *Proceeding of RECITAL(Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues).* Avignon. 9-13 join 2008. http://www.lia.univ-avignon.fr/index.php?id=644.

Busemann, S. Decleck. T.. Diagne. A. K.. Dini. L.. Klein. J. and Schmeier. S. 1997. “Natural Language Dialogue Service for Appointment Scheduling Agents.” *Proceedings of the Fifth Conference on Applied Natural Language Processing,* pp.25-32.

Ferro, L. Gerber, L. Mani, I. Sundheim. B. & Wilson G. 2005. *TIDES 2005 Standard for the Annotation of Temporal Expressions.* April 2005 the September version updates all references to the ISO 8601 standard to incorporate the latest edition (8601:2004. Third Edition). http://fofoca.mitre.org/annotation_guidelines/timex2_annotation_guidelines.html

Gross, M. 2002. Les déterminants numéraux. un exemple : les dates horaires. Langages 145, pp.21-38.

ISO-TimeML Working Group. 2008. ISO/DIS 24617-1. *Language Resource Management: Semantic Annotation Framework Part1: Time and Events.* ISO/TC37/SC4 (published on Oct. 2008). 151p.

Saurí, R., Littman, J., Knippen, B., Gaizauskas, R., Setzer, A. & Pustejovsky, J. 2006. *TimeML Annotation Guidelines Version 1.2.1.* 71p.

Verhagen, M., Mani, I., Saurí, R., Knippen, R., Littman, J. & Pusterjovsky, J. 2005. Automating Temporal Annotation with TARSQI. *In Proceedings of the ACL 2005*.

1차 자료

정지영. 홍재성 편저. 1998. 『프라임 불한 사전』. 서울: 두산동아.

한국불어불문학회. 2007. 『새한불 사전』. 서울:한국외국어대학교출판부

Arrivé, M., Gadet, F., Galmiche, M. 1986. *La grammaire d’aujourd’hui.* Paris: Flammarion.

Charaudeau, P. 1992. *Grammaire du sens et de l’expression.* Paris: Hachette Education.

Grand Robert. *Le Grand Robert: Dictionnaire Alphabétique et Analogique de la langue française.* (version en CD-ROM)

Grevisse, M. 1980. *Le bon usage: Grammaire française avec des remarques sur la langue française d’aujourd’hui.* Paris: Duculot.

Mauger, G. 1968. *Grammaire pratique du français d’aujourd’hui.* Paris: Hachette.

TLF. *Trésor de la langue française Informatisé.* http://atilf.atilf.fr/dendien/scripts/tlfiv4/showps.exe?p=combi.htm;java=no;

윤애선
부산광역시 금정구 장전동 산 30번지
부산대학교 불어불문학과
609-735
E-mail: asyoon@pusan.ac.kr

접수일자: 2008. 11. 18
수정일자: 2008. 12. 05
게재일자: 2008. 12. 07