현대 한국어에서 아라비안 숫자의 읽기 규칙 연구

정영임†, 김정세‡, 김상훈‡, 이영직‡, 윤애선†
†부산대학교 인지과학 협동과정, ‡한국전자통신연구원
{acorn, asyoon}†@pusan.ac.kr, {jungskim, ksh, ylee}‡@etri.re.kr

A Study on the Arabic numeral reading rules in Modern Korean

Youngim Jung†, Jeongse Kim‡, Sanghoon Kim‡, Youngjik Lee‡, Aesun Yoon†
†Dept. of Cognitive Science, Pusan National University,
‡Spoken Language processing Team, ETRI

요 약

본 논문에서는 아라비안 숫자를 포함한 텍스트를 음성으로 합성하기 위하여, 숫자 형태와 분류사 그리고 숫자가 나오는 문맥에 따라 숫자를 자동으로 문자화할 수 있는 전처리 규칙을 설정하는데 목적을 둔다. 먼저 선행연구를 통해 숫자를 포함한 수사 및 수사표현의 읽기 규칙의 적용 범위 및 한계점을 살펴 보고, 음성 합성을 위한 아라비안 숫자의 문자화 규칙을 설정하고자 한다. 현대 한국어에서 아라비안 숫자를 읽는 방식은 크게 고유어 방식과 한자어 방식이 있으며, 단(單)단위에서는 영어가 사용되기도 한다. 또한 한자어 방식에서도 단위를 붙여 읽는 경우와 모든 수를 단 단위로 읽는 경우가 있으므로, 아라비안 숫자의 문자화를 단순한 규칙을 설정하여 자동화하기에는 중의성이 높다. 본 연구에서는 ① 숫자 전치어(pre-numeral), ② 기호를 포함한 숫자열의 표현 형식과 크기, ③ 단위 표현, ④ 숫자 후치어(post-numeral), ⑤ 분류사(classifier), ⑥ 분류사 후치어(post-classifier), ⑦ 수사표현 앞뒤 문맥에 따라, 아라비안 숫자 표현이 문자화되는 방식을 살펴보았다. 분석 대상 말뭉치는 C 신문의 2000년 1월부터 2000년 4월까지 전체 기사 1,400건에서 숫자가 포함된 숫자표현 약 63,000개로 구성하였다. 패턴화된 구조 및 중의성이 없는 구조를 12가지로 밝히고 중의성이 있는 구조의 유형을 밝혔으며 분류사 후치어와의 결합 관계, 좌우 문맥정보를 통해 중의성 해결의 단서를 제시하고자 하였다.

1. 서론

정보화 사회로의 진입이 시작한 지 불과 5-6년 만에 컴퓨터, 네트워크, 휴대폰, PDA 등은 현대인의 일상생활에 없어서는 안 될 도구로 평가받고 있고, 언제 어느 때나 정보의 접속과 처리가 가능하도록 진행되고 있다. 따라서 정보기기는 통합화를 지향하고, 휴대가 가능하도록 급속도로 경량화 및 무선화되어 간다. 정보기기를 경량화하기 위해 현재 문자(text) 중심의 입출력 장치를 음성(voice) 중심으로 변화시키고자, 음성합성(text-to-speech)과 음성인식(voice recognition)에 대한 다양한 연구가 활발히 진행되고 있다. 음성 합성의 경우, 아직 문장이나 구(phrase) 단위의 어조(intonation)가 어색하고 연음처리가 부자연스러우나, 음절이나 단어 단위의 합성음은 정확하게 생성되므로, 실생활에서 전화번호나 증권정보 안내 등에 사용되고 있다. 한국어 음성 합성을 위해서는 여러 단계의 처리가 필요하나 우선 숫자, 기호, 이니셜, 외국어, 수식 등이 포함된 문서를 한국어 음성 생성 규칙이나 휴리스틱스 등이 적용될 수 있는 문자로 전처리(pre-processing)해주는 작업이 필요하다. 문서의 종류에 따라 나타나는 비-문자의 종류와 비율이 다르나, 정확한 정보 전달을 목적으로 하는 문서일수록 가독성(readability)을 높이는 아라비안 숫자의 사용 비율이 높다.

한국어에서 아라비안 숫자(이하 '숫자'로 표기함)를 읽을 때 '하나, 둘, 셋'과 같이 고유어를 사용하기도 하고 '일, 이, 삼'처럼 한자어로 읽기도 한다. '2457'과 같은 숫자를 한자어를 사용해 읽을 때, 문맥에 따라 한편으로는 단위를 추가하여 '이천사백오십칠'로 읽는가 하면, 다른 한편으로는 단위 없이 '이사오칠'과 같이 읽기도 한다. 그리고 서구 문화의 영향으로 숫자도 영어와 같은 외래어로 읽는 경우도 발생한다. 특히 신조어나 외래어가 신속하게 발생하고 소멸되는 현대 한국어에서 숫자를 읽는 방식이 매우 다양하며 중의성이

높으나, 이에 대한 연구는 거의 이루어지지 않았다.

이에,. 본 연구에서는 정제된 언어를 사용하지만 현대 한국어의 변화상을 살펴볼 수 있는 신문 자료를 대상으로, 음성 합성을 위한 숫자의 문자화를 위한 전처리 규칙을 설정하는데 그 목적이 있다.[1] 대상 말뭉치로는 C 중앙지의 기사 1,400건에서 숫자가 포함된 숫자표현 약 63,000개를 분석하였다.[2] 이 자료의 분석을 바탕으로 패턴화된 숫자표현을 찾아내고, 숫자표현의 구성성분 간 결합 관계를 분석하여 중의성이 낮은 아라비안 숫자의 문자화 규칙을 설정하고자 한다.[3]

2장에서는 숫자의 문자화에 관련된 선행 연구를 살펴보고 그 한계점을 알아본다. 3장에서는 숫자열의 앞뒤어절로 구성된 '수사표현'의[4] 구성요소 및 구성요소 간 띄어쓰기 규칙을 알아본다. 4장에서는 숫자의 전후 어절을 포함한 정보를 분석하여 숫자의 문자화 규칙을 설정하며, 5장에서는 남은 문제 및 향후 연구 방향을 살펴본다.

2. 선행 연구

본 연구의 내용과 관련된 선행 연구는 많지 않으나, 학문 분야에 따라 크게 3가지 관점으로 분류할 수 있다. 첫째, 이론 국어학에서는 수사의 품사 및 형태 분석과 함께, 의존 명사로서 분류사의 종류에 따른 수사의 선택이 주로 논의되었고,[1, 11, 12, 13, 14] 둘째, 정보 처리에서는 정보 추출이나 요약에 사용될 수 있도록 수사 표현의 구성 요소를 분석하고 그 결합 관계를 밝히고자 하였으며,[4, 5] 셋째, 전산 언어학이나 음성 처리 분야에서 숫자 읽기 규칙의 일부가 분석 및 기술되었다.[7, 8, 9] 그 밖에 실용적 연구로 방송 언어 연구분야에서 숫자의 바른 읽기 방식을 제시한 바 있다.[2, 3] 본 연구의 목적인 '숫자의 읽기 규칙'을 설정하기 위해 선행 연구에서 제시한 결과와 그 한계점을 다음과 같이 요약할 수 있다.

2.1. 수사의 종류

수사는 크게 ㉠ 언어 계통에 따라서 순수 한국어인 고유어 계통(예: 일곱, 열)과 한자어 계통(예: 칠, 십)로 구분하며, 두 계통 모두 ㉡ 서수(예: 첫, (제) 일)와 기수(예: 한, 일) 표현을 가지고, ㉢ 정(定)수(예: 한, 두, 일, 이)와 부정(否定)수(예: 한두, 서너, 댓, 일이)를 나타내는 표현을 갖는다. ㉣ 기수의 기본형(예: 세, 네, 다섯, 여섯, 육, 십)은 특정한 분류사와 결합 관계에서 변이형(예: 서, 석, 너, 넉, 닷, 엿, 유, 시)을 갖는다.[5] ㉤ 고유어 계통은 품사에 따라서 명사(예: 하나, 둘)와 관형어(예: 한, 두)로 구분된다. 한자어 계통은 ㉥ 수의 형태에 따라 정수(整數), 분수(分數), 소수(小數)를 모두 나타내며, ㉦ 정수와 소수의 경우, '점, 십, 백, 천, 만' 등 자릿수를 표시하는 단위와 사용되는 경우와 자릿수 단위없이 사용되는 경우로 분류한다. 선행 연구에서 수사의 종류에 대해 비교적 세분화된 분류를 제시하였으나, 실제 현대 한국어에 나타나는 영어 수사(예: 투 아웃)를 포함하지 못 했다는 한계를 갖는다.

2.2. 분류사와 수사의 선택 제약

분류사와 수사의 선택 제약 관계에 대해 채완(1983)은 "분류사가 고유어 계통이면 고유어 수사를 선택하며(예: 서 되, 다섯 마리), 분류사가 한자어이거나 외래어면 한자어 수사를 사용한다(예: 삼년(年), 오미터(m))"는 원칙을 제시하고 있다. ㉠ '명(名), 시(時), 개(個), 살, 달, 시간(時間), 군데, 마리, 해, 가지, 사람' 등은 고유어 수사와 결합하고, ㉡ '원, 년(年), 일(日), 세(歲), 월(月), 도(度), 퍼센트(%), 개월(個月)' 등은 한자어 수사와 결합한다. 고유어 수사와 한자어 수사 모두 결합하는 종류로는 ㉢ '분(분 vs 分), 대(대, 臺 vs. 代), 기(基 vs. 期)'와 같은 동형이의어(예: 러시아 군사학교 삼기 졸업생은 미사일 세 기를 이라크에 팔았다.), ㉣ '동, 층, 페이지' 등과 같은 서수와 기수의 차이(예: 아파트 일곱 동 중 칠동의 지하 주차장이 붕괴되었다.), ㉤ 의미의 차이 없이 고유어 수사나 한자어 수사가 모두 사용되는 경우(예: 서른두 평 아파트, 삼십이 평 아파트)로 구분한다.

분류사와 수사 간의 선택 제약에 관해서는 후자처럼 목록을 제시하는 것이 언어 현상을 좀더 타당한 분석 방법이다. 하지만 선행 연구에서 목록의 분류가 정확

---

[1] 전화번호, 차번호, 우편번호, 계좌번호와 같이 기호를 포함한 일련의 숫자는 본 연구의 범위에 포함되나, '사과 세 개, 三國誌, 한국쓰리엠'와 같이 수사를 한글로 적거나 한자로 적어 놓은 것, 또는 외래어 발음을 한글로 표기한 것과 같이 표기한 대로 발음하는 수사는 제외된다.

[2] Random Sampling 방식으로 추출된 말뭉치는 C 신문에서 제공하는 전자 아카이브에서 2000년1월1일에서 2000년 4월 30일까지 저장된 전체 기사로 구성된다. 숫자의 출현 비율은 정치면이나 사회면보다는 경제면, 스포츠면 기사에서 월등히 높다.

[3] 'Yes24.com(예스이십사닷콤)', 'CJ39쇼핑(씨제이삼구쇼핑)', 'SM5(에스엠파이브)', 'SM520(에스엠오이공)'과 같이 숫자를 포함한 고유명사의 경우, 숫자의 발음에 일괄적으로 적용할 수 있는 기준이 없다. 이에 본 연구에서 고유명사에 나타나는 숫자의 정형화는 고려하지 않았다.

[4] 본 고에서 사용되는 '수'와 관련된 용어의 구분은 다음과 같다. '숫자'는 '아라비안 숫자'의 약칭으로 사용하며, '수사'는 아라비안 숫자를 비롯하여 문자로 표시된 수 표시 형태를 포함하는 좀 더 포괄적인 용어로 사용하고, '수사 표현'은 수사와 통사적, 의미적 결합관계를 가지며 수량 표시에 사용되는 전체 언어 단위를 지칭하는데 사용한다.

[5] '서, 석, 너, 넉'은 고유어 분류사 '되' 등과 결합하는 '세, 네'의 변이형이며, '유, 시'는 '월(月)' 앞에서만 나타나는 '육, 십'의 변이형이다.

하지 않은 경우가 다수 존재하며[6], 목록이 완전(exhaustive)하지 않다는 한계점을 갖는다. 특히 고유어 수사와 한자어 수사를 모두 선택할 수 있는 분류사에 대해, 수사표현을 구성하는 다른 요소나 좌우 문맥을 이용하여 분류사의 중의성을 해결하려는 시도는 전혀 이루어지지 않았다.

2.3. 수의 크기와 수사의 선택

고유어 수사와 결합하는 분류사의 경우, 숫자가 커지거나 숫자의 형태에 따라 한자어 수사를 함께 사용하거나 오히려 한자어 수사가 더 자연스럽다는 관찰은 여러 선행 연구에서 이루어졌다. 하지만 경계가 되는 수의 크기에 대해서는 몇 가지 다른 기준이 제시되었다. 채완(1983)은 10을 기준으로 하여 미만과 이상일 때 각각 고유어와 한자어를 더 선호하는 현상을 지적하고, 100 이상에서도 1단위가 1~9일 때 고유어 수사가 더 자연스럽게 사용된다고 한다. 반면에 유재원(1999)은 그 기준을 20과 100으로 세분화하여 20 미만과 100이상에서는 각각 고유어 수사와 한자어 수사가 배타적으로 사용되며, 그 사이에는 한자어 수사와 고유어 수사의 사용이 모두 허용된다고 한다.

하지만 언어 규범의 관점에서는 이러한 수의 크기에 따른 수사의 교체 현상을 인정하지 않아[2, 3] 고유어 수사와 한자어 수사의 선호도는 개인의 차가 매우 커 일반화된 규칙으로 설정하기 힘들다.

2.4. 수사 표현의 패턴과 수사와의 관계

수사 표현의 패턴이나 문맥이 수사의 읽기 방식에 영향을 준다는 점은 김상준(1986, 1992)에서 예시를 통해 제시되었다.[7] '-'나 '~'와 같은 기호가 포함된 숫자열로 전화번호, 주소의 번지, 차량번호 등의 패턴 읽는 방법을 상세히 기술하고, '.' 기호가 포함된 있는 숫자열에서 소수점 이하 자리에 나타나는 숫자 '0'이 '영'과 '공'으로 읽히는 규칙을 제시하였다. 이와 함께 기념일을 읽는 예를 보여준다.

하지만 선행 연구에 나타난 패턴의 수가 실제의 경우보다 많지 않다. 예를 들어 전화번호, IP주소, 아파트 동·호수 등이 더 세분화된 패턴을 보이며, 주민등록번호, 우편번호 등도 일정한 패턴을 가지므로 이에 대한 기술이 필요하다.

2.5. 수사 표현의 구성 요소와 띄어쓰기

국어학에서의 연구를 통해 수사 표현의 구성 요소와 구성 요소 간의 결합 관계가 제시되었는데 특히 수량사구 유형을 명사, 수사, 분류사의 어순과 분류사의 유무에 따라 네 가지로 나누고 각각의 유형에 선택되는 수사를 설명하였다.[11] 그러나 수사 앞에 오는 요소로 접두사나 기호 및 부호를 고려하지 않았고, 분류사 후치어에 대한 언급이 없어 이를 자동처리를 위한 숫자 표현의 결합관계에 적용하기에 분석이 완전하지 못하다.

또 채완(1983)은 '명사+수사'의 구조에서 수사는 '고유어 수사의 명사형'만이 가능하다고 지적하였으나 실제 신문 데이터에 나타나는 형태를 살펴보면, '명사 + 한자어 수사'가 일반적이었다. (예: '승점 3으로 이길 수 있었던 삼성', '그림 ②를 참조')

자동처리에 중요한 요소인 수사 및 숫자 표현의 띄어쓰기에 대해서는 선행연구된 예가 발견되지 않는다. 다만 한글 맞춤법 어문 규정과 '한글 전용 편람'에 수사구 띄어쓰기에 대한 규정이 있지만 상세하지 않고 '어울려 쓸 수 있는 경우 붙여쓴다.'와 같이 규정이 모호하다.

3. 수사표현의 구성요소

현대 한국어에서 아라비안 숫자는 분류사, 수의 크기, 패턴, 구성요소 간 결합관계, 문맥 등에 따라 다양하게 읽어야 한다. 그러나 2장에서 살펴본 바와 같이 선행 연구에서 제시된 수사구조로는 실제 신문데이터의 다양하고 중의적인 수사 및 숫자 표현을 분석하는데 적절하지 못하다. 3장에서는 수사를 분류하고, 숫자를 포함한 수사 표현의 구성요소를 살펴 보며, 구성요소 간 띄어쓰기 규칙을 알아본다.

3.1. 숫자 읽기 방식의 분류

선행 연구의 결과와 본 연구에서 구성한 말뭉치에 기초하여, 문자화된 숫자를 ㉠ 품사, ㉡ 언어 계통, ㉢ 기수/서수, ㉣ 정수/부정수, ㉤ 자릿수를 나타내는 단위의 유무, ㉥ 기본형/변이형에 따라 <표1, 2, 3>과 같이 구분할 수 있다.[8]

<표1: 명사 수사의 분류>

구분			약어	예
고유어	기수	정수	Kca_n	셋, 넷
		부정수	Kca_ni	서넛
	서수		Kor_b	셋째
한자어	기수		Cca_b [+u]	삼, 사
	서수		Cor	(제)삼

[6] 예를 들어 '세기(世紀)'와 분류사를 한자어 수사와 결합하는 방식으로 분류하였으나, 기수의 경우에도 '한 세기'와 '일 세기' 간에는 의미 차이가 없다.

[7] 채완(1983)은 수학 계산식 읽기 방식을 보여주고, 이영직(2000)은 분류사가 결합되지 않는 유형으로 구분하여, 스포츠 경기 스코어, 분수, 지진의 세기, 시력, 전화번호, 특수 날짜 읽는 방식의 특수성을 간략히 지적하였다.

[8] <표1>, <표2>의 'Kca_n' 등의 영문 기호는 4장에서 규칙을 기술할 때 수사의 종류를 명시하기 위한 약어이다.

<표2: 관형어 고유어 및 영어 수사의 분류>

구분				약어	예
고유어	기수	정수	기본형	Kca_b	세, 네
			변이형	Kca_v	서(석), 너(넉)
		부정수		Kca_i	서너
	서수	정수		Kor_b	셋째, 넷째
		부정수		Kor_i	서너째
영어				Brn	원, 투

<표3: 관형어 한자어 수사의 구분>

구분					약어	예
기수	정수	+단위	정수	기본형	Cca_b [+U]	육, 십 이백십오
				변이형	Cca_v [+U]	유, 시
			부정수		Cca_i	삼사, 삼사십
		-단위			Cca_b [-U]	이일오
	소수	+단위			Cca_d [+U]	이점일오
		-단위			Cca_d [-U]	이일오
	분수				Cca_f	이와 오분의 일
서수	정수				Cor_b	(제) 일
	부정수				Cor_i	(제) 삼사

3.2. 수사표현의 결합구조

3.1에서 살펴본 것처럼 아라비안 숫자는 문자화하는데 중의성이 매우 높아 자동 변환 규칙을 설정하기 힘들다. 따라서 숫자와 의미 통사적으로 밀접한 관계를 맺는 수사표현의 각 구성 요소와의 선택 관계는 중의성을 낮추는데 매우 큰 역할을 한다. 관형어 수사 및 명사 수사가 출현하는 수사 표현의 결합구조는 <표 4, 5>와 같다.

<표4: 관형어 수사의 수사표현의 결합구조>

수사표현 좌문맥		
수사 표현 구조	숫자 전치어	①
	기호포함 숫자열	②
	문자 단위표현	③
	숫자 후치어	④
	분류사	⑤
	분류사 후치어	⑥
수사표현 우문맥		

① 숫자 전치어(pre-numeral)의 종류로는 ⓐ '₩(원), ¥(엔), ₱(페소), €(유로), £(파운드), $(달러)" 등 통화 기호, ⓑ '마하'와 같은 측정 단위, ⓒ '제'와 같은 서수 표현, ⓓ '총(總), 총액(總額), 약(略), 만(滿)', ⓔ '새벽, 오전, 낮, 오후, 밤, 상오, 하오, 현지 시각' 등 시간 표현, ⓕ '승점(勝點)'과 같은 스포츠 용어, '지수(指數)'와 같은 주식 용어, ⓖ '여의도동, 을지로1가' 등 주소 표현, ⓗ 차량 번호를 구성하는 '서울, 부산, 경남' 및 '가, 모, 더'를 들 수 있다.

② 숫자열에 포함된 기호로는, ⓐ'+(플러스/더하기), -(마이너스/빼기), =(은/는), :(대)'와 같은 수학 기호, ⓑ '번지, 의(에), 에서'로 읽히거나 문자화되지 않는 '-' 또는 '~', ⓒ 시간 표현에 사용되는 ':', ⓓ ','나 소수 표현에 사용되는 '.' 등을 들 수 있다.

흔히 신문 자료에서 '3,200'이나 '2,350,978,000' 등을 표기할 때, 아라비안 숫자열을 길게 나열하는 것보다 '3천2백, 23억 5,097만 8천'과 같이 문자열을 섞은 표현 형식을 더 많이 사용한다. 이때 사용되는 ③ 문자 단위표현은 '십, 백, 천, 만, 억, 조, 경, 해' 등이다.

④ 숫자 후치어(post-numeral)로는 '그 정도나 그 이상'의 의미를 나타내는 '여(餘)'가 있다.

⑤ 분류사(classifier)의 어원은 숫자열 선택에 중요한 기준이 되는데 그 종류로는 ⓐ '자루, 말, 사람, 채, 대, 도움[9]'등과 같은 고유어 계통, ⓑ '개(個), 명(名), 대(代, 大, 臺), 위(位)'나 '개국(個國), 개소(個所), 구원승(救援勝)'과 같은 한자어 계통, ⓒ '미터(meter), 골(goal), 파(par)'와 같은 외래어 계통을 들 수 있다.

⑥ 분류사 후치어(post-classifier)로는 '경(頃), 선(線), 대(臺), 째, 씩, 어치, 짜리, 쯤, 여(餘)' 및 '이상, 이하, 미만, 초과, 가량, 남짓, 정도' 등을 들 수 있다.

수사 표현 좌문맥과 우문맥은 각 숫자 표현에 따라 매우 다양하게 나타나므로 일반화하기 힘들다. 그러나 중의성이 있는 숫자의 의미와 발음을 파악할 때, 수사표현 좌우문맥이 중요한 단서를 제공할 수 있다. 예를 들어 좌문맥으로는 '우편번호, 전화(번호), 팩스(번호), 차량번호'나 '요한복음' 등을, 우문맥으로는 '3번 지방도로', '5번 시드' 등에서의 '지방도로', '시드'와 '3장을 세어라.'(세/*삼 장을 세어라), 3장 24절(*세/삼 장, 이십사 절)'에서 '세어라'나 '24절'과 같은 공기(coocurrence) 정보나 연어(collocation) 관계 등을 들 수 있다.

<표 5: 명사 수사의 수사표현의 결합구조>

수사표현 좌문맥		
수사표현 구조	숫자 전치어	n①
	기호포함 숫자열	n②
	숫자 후치어	n③
수사표현 우문맥		

명사 수사가 포함된 수사 구조에서 n① 숫자전치어로는 ⓐ 사람을 나타내는 명사(예: 사람, 학생, 남자, 여

[9] '도움, 튄공' 등은 'assist, rebound' 등의 외국어 스포츠 용어의 순화 용어이다.

자, 소녀, 간호원, 교사, 군인 등)나 ⓑ 목록에 포함될 수 있는 '사과', '배', '공책', '책상' 등의 가산명사[10], ⓒ '그림', '표', '그래프', '차트' 등과 ⓓ 전문용어인 '승점', '흑', '마하'등이 있다. n② 기호포함 숫자열에는 다단계 번호를 나타내는 부호 '-', '.' 등이 포함될 수 있다. (예: 그림 2-2) n③ 숫자 후치어로는 각 종 조사 및 '-이다' 어미가 올 수 있다.

3.3 수사표현의 구성성분 간 띄어쓰기 규칙

수사표현이 여러 개의 구성성분으로 이루어지므로 한 어절에 나타나기보다는 여러 어절에 걸쳐 나타나는 경우가 더 많다. 나타나는 어절수가 늘어날수록 검색 시간도 현저히 길어지고 검색 속도가 떨어진다. 한글 맞춤법 어문규정[11], 에 의하면 수사표현의 구성요소 간 띄어쓰기 규칙을 다음과 같이 제시하고 있다.

(a) ①이 ②의 접두어일 경우는 붙여쓰고 관형어일 때 띄어쓴다.(예: 제7차 회담, 금 15일) 통화단위(예: ₩, ¥,…)와 숫자 사이의 띄어쓰기는 규정에 없다.
(b) ②와 ③는 붙여쓴다. 단, '②+③'의 구조가 2번 이상 반복되는 경우 '만, 억, 조, 경' 등 4자리 단위로 띄어쓴다. (예: 12억 3456만 7869/ 십이억 삼천사백오십육만 칠천팔백구십팔)
(c) ②와 ④는 붙여쓴다. (예: 10여 일 간, 십여 호)
(d) ②와 ⑤는 붙여쓴다. 순서를 나타내는 경우나 숫자와 어울리게 쓰이는 경우에는 붙여쓸 수 있다. (예: 16동 502호, 10개)
(e) ③과 ④는 붙여쓴다. (예: 2천여 호, 오만여 개)
(f) ③과 ⑤ 사이는 띄어쓴다. (예: 255만 가구)
(g) ④와 ⑤ 사이는 띄어쓴다. (예: 십여 호)
(h) ⑤와 ⑥ 사이는 붙여쓴다. 단, 접미사가 아닌 분류사 후치어(예: 이상, 이하, 미만, 초과, 가량, 남짓)는 띄어쓴다. (예: 10개째, 열 사람 가량, 4배 이상)

하지만 실제 신문 자료에서는 이러한 띄어쓰기 규칙 적용이 임의적이다. 한 신문 자료에 나오는 띄어쓰기라도 규칙에 맞게 표기하기도 하였으나 아래의 예와 같이 어긋나기도 하였다.

(a)' 1962년 제 1차 경제개발 계획
(b)' 2만6천3백1
(e)' 10만 여명의
(f)' 1천9백~2천 가구
(g)' 216만여명
(h)' 5000만달러 어치, 12시간이상

4. 숫자 읽기 규칙

3장에서 분석된 수사 표현의 구성요소를 바탕으로 4장에서는 각 구성요소의 결합에 따라 패턴화된 결합 구조, 중의성이 없는 결합 구조에서 숫자 읽기 규칙을 설정한다. 그러나 규칙을 통해 숫자 읽기를 해결할 수 없는 경우가 많은데, 이는 수사의 크기에 따라 교차수사를 허용하는 분류사, 기·서수 의미에 따라 동형이의어인 분류사가 있으며 실제 신문데이터에서 일정한 규칙없이 사용되는 기호에 의해 발생하는 중의성때문이다. 숫자의 음성합성을 보다 완전한 목록화를 통해 이루어질 수 있도록 중의성이 있는 결합 구조를 밝히고 이를 해결할 수 있는 방법을 제시하였다.

4.1 패턴화된 결합 구조

숫자 표현 중 일정한 자리수의 숫자가 오거나 숫자 사이에 일정한 규칙으로 기호가 들어가 사회적으로 통용되는 표현의 구조는 정형화되어 있다. 예를 들어 전화번호, 차량번호, 주민등록번호 등은 아래와 같이 정형화된 구조를 가지고 있으며[12] 수사의 의미보다는 기호의 의미가 강해 숫자의 발음에 있어서도 정형성을 가진다. 이러한 정보는 별도로 분류할 경우 컴퓨터 기반 수사 처리의 효율성을 높이는데 도움을 준다.

(a) 전화번호: 유선전화번호와 휴대전화번호에 따라 두 가지 패턴화된 구조를 가진다.

ⓐ 유선전화번호

NNA	-	LNA	-	?	?	N	N	-	?	N	N	N

ⓑ휴대전화번호

NNA	-	SNA	-	?	N	N	N	-	N	N	N	N

전화번호에 나오는 번호는 모두 Cca_b[-U]로 읽되 국번호와 고유번호는 단위를 넣어(Cca_b[+U]) 읽기도 한다. 국가번호(NNA), 지역번호(LNA) 및 서비스사번호(SNA)는 리스트될 수 있다.

(b) 차량 등록번호:

LST	?	N	C	N	N	N	N

지역을 나타내는 문자열(LST)과 숫자의 결합인 차량번호는 Cca_b[+U]와 Cca_b[-U]의 두 가지로 읽을 수 있다.

(c) 주민등록번호:

N	N	N	N	N	N	-	N	N	N	N	N	N	N

[10] 채완(1983)에 의하면 사람을 나타내는 명사(예: 남자 셋, 학생 열 다섯)가 올 때 가장 자연스럽고 불가산명사는 올 수 없다고 하였다.

[11] 여러 띄어쓰기 규칙이 모호하듯이, 수사표현 구성성분간 띄어쓰기 규칙도 거의 정의되지 않거나 예외 조항이 많다. 3.3은 1988년도에 개정된 한글 맞춤법 어문규정(제5장 띄어쓰기 제2절 제43항)의 내용을 근간으로 하되, 1969년 개정된 '한글 전용 편람'에서의 띄어쓰기 규정 및 (국정 교과서에서의) 붙여쓰기 용례를 참고하였다.

[12] 'N'은 필수적으로 요구되는 숫자, '?'는 선택적 수사, 'C'는 문자를 의미한다. 'N'사이의 '…'는 'N'의 개수가 정해지지 않았음을 의미한다.

Cca_b[-U]로 읽는다.

(d) IP 주소:

?	?	N	.	?	?	N	.	?	?	N	.	?	?	N

Cca_b[-U]로 읽는다.

(e) 시각:

TST	?	N	:	N	N	:	?	?

':'을 '시', '분', '초'로 바꾸어 읽되 '시' 앞의 숫자는 Kca_b로 읽고 '분', '초' 앞의 숫자는 Cca_b[+U]로 읽는다. '오전', '오후', '새벽' 등은 시간대를 나타내는 표현(TST)이다.

(f) 특수기호를 전·후치한 숫자 표현

ⓐ 통화기호(CS)를 포함한 통화표현

CS	N	N	,	N	N	N	,…	,	N	N	N

숫자는 Cca_b[+U]로 읽되 통화기호(CS)를 숫자 뒤로 붙여 읽는다.(예: ¥100(백 엔))

ⓑ 온도

?	?	?	N	℃/℉

℃나 ℉는 각각 섭씨와 화씨로 고쳐 숫자 앞에 붙여 읽는다. (예: 27℃(섭씨 이십칠 도)) 이외에도 법인번호, 성경의 장·절 구조, 주소 표현 및 우편번호, 연호가 있는 연도 등의 패턴화가 가능하다.

4.2. 중의성이 없는 결합 구조

숫자의 형태, 숫자 전치어, 문자 단위표현, 숫자 후치어 및 분류사, 분류사 후치어에 따라 숫자 표현의 의미와 발음이 결정되는 경우가 있다. 이러한 구조 역시 '패턴화된 결합 구조'처럼 범주화가 가능하다. 숫자 표현이 아래의 조건 중 어느 하나라도 만족되면 중의성이 없는 결합 구조를 이룰 수 있다.

(a) '② 기호포함 숫자열'의 숫자 형태가 분수의 경우 고유한 발음을 가지며,(예: ⅞(팔 분의 오), 3⅝(삼과 팔 분의 오)) 신문지상에는 문자코드때문에 '3과 8분의 5'처럼 표기하기도 한다. 또한, 지수를 포함한 숫자의 경우, 밑수는 한자어 읽기하고 지수는 Kca_b+'제곱'으로 읽는다.(예:5^4 (오의 네제곱))

(b) '①숫자 전치어 + ②기호포함 숫자열'의 구조에서 ① 중 '제', 스포츠 용어 '승점', 주식 용어 '지수', 전문 용어 '마하' 등이 올 때 ②는 Cca_b[+U]이다. (예: 제7차 장관급 회담) 그리고, ① 중 '+', '-'가 오면 ② 앞에 각각 '플러스', '마이너스'를 붙여 Cca_b[+U]로 읽는다. (예: -130%(마이너스 백삼십 퍼센트)

(c) '② 기호포함 숫자열 + ③ 문자 단위표현'의 결합 구조에서 ②는 항상 Cca_b[+U]이다. ②가 '1'일 때는 '일'의 발음을 생략하나 '억', '조' 등의 앞에서는 '일'을 생략하지 않는다. (예: 1억1만1천1백50(일억 만 천백오십))

(d) '②기호포함 숫자열 + ④ 숫자 후치어(餘)'의 구조는 ②가 단위수(단단위가 0인 수)일 때만 가능하고 Cca_b[+U]로 읽는다. 예) 10여 개(십 여 개)

(e) '② 기호포함 숫자열 + ⑤ 분류사'에서 ⑤가 서구로부터 들어온 도량형 단위(예: 3m, 3㎭)이거나 한자어 분류사 중 순서를 나타내는 분류사(예: 1위(位), 3차(次), 5학년(學年)) 및 서구에서 차용된 스포츠 용어 중 '오버파', '온', '구원승'과 같은 분류사는 Cca_b[+U]만 결합한다. 각 분류사의 개수는 아래와 같다.[13]

<표 6: Cca_b[+U]만 결합하는 분류사>

서구도량형	126 개
순서 한자어 분류사	91 개
스포츠용어 분류사	21 개

(f) 그 외에 ① 중 새벽, 오전, 낮, 오후, 밤, 상오, 하오가 오고 ⑤ 중 '시', '분', '초'가 올 경우, '시' 앞의 수사만 Kca_b로 읽고, '분', '초' 앞의 수사는 Cca_b[+U]로 읽는다. (예: 오후 5시 25분(다섯 시 이십오 분))[14].

4.3. 중의성이 있는 결합 구조

비록 앞 절에서 패턴화와 중의성이 없는 규칙이 제대로 적용된다하더라도, 모든 언어에는 한정된 표현 수단으로 무한한 의미를 나타내야하므로 의미의 중의성을 피할 수 없는 표현이 많다. 숫자 표현에서도 동일한 기호를 여러 가지 용도로 쓰거나 동형이의어 분류사가 광범위하게 분포하며 때로는 동형의 조사와 결합하여 중의성이 발생한다. 또한 기호를 잘못 표기하여 중의성이 생긴다. 중의성이 있는 결합 구조는 ㉠숫자 표현 구성 요소 ㉡좌우 문맥 요소를 살피며 필요한 경우 ㉢자동처리 전(前)단계에서 중의성을 발생하는 오류 수정작업을 통해 중의성을 줄일 수 있다.

(a) '②기호포함 문자열': 숫자 사이에 기호를 넣어 여러 가지 의미로 사용하고 있다. '-' 기호는 ㉠부정수(예: 2-3개(두세 개)), ㉡범위수(예: 5-7일(오일에서 칠 일)), ㉢스포츠 경기에서 점수(예: 2-3(이 대 삼))나 토너먼트 경기의 명칭(예: 3-4위전(삼사위전)), ㉣ 수학식의 일부분(예: 3-2(삼 빼기 이)) 등에 사용된다.

'~' 기호는 ㉠부정수(예: 2~3개(두세 개)), ㉡범위수(예: 5~7일(오 일에서 칠 일))나 IP의 하위 디렉토리를 표시할 때 사용되고 있다.

---

[13] 서구 분류사는 [14]를 참고하였고 순서한자어와 스포츠 용어 분류사는 대상 신문데이터를 통해 수집되었다.

[14] '14시(십사 시)'는 군대식 표현이므로 사용하지 않는다.

':'의 경우, 시간대를 나타내는 표현(TST)이 없을 경우, 둘 이상의 대비에 사용되고(예: 2:3(이 대 삼)) '/' 기호는 ㉠분수를 표시하거나(예: 2/3(삼분의 이)) ㉡날짜 표현에 사용되는데(예: 2/3(이월 삼일)) 년도까지 함께 표현되거나(예:2002/2/3(이천이년 이월 삼일)), '일자', '날짜', '갱신일', '저장일'과 같은 단어가 오면 날짜 표현이다.

','는 ㉠ 숫자를 열거할 때(예: 2, 3, 4, 5(이, 삼, 사, 오)), ㉡ 부정수(예: 2, 3개(두세 개)), ㉢수의 자릿점을 나타낼 때(예: 14,314) 사용된다. 숫자의 열거와 부정수 표현에 사용되는 ','는 숫자 사이를 띄나 수의 자릿점으로 쓰일 때는 숫자 사이에 공백이 없다. 부정수는 숫자 2개를 연결하고, 숫자 열거는 숫자가 2개 이상이며 분류사가 후치하지 않는다.

'.'은 숫자와 관련해서는 소수를 표현할 때만 사용된다. 그러나 전자화된 기사에서는 코드 사용의 문제로 기념일 표시에서 숫자 사이의 가운뎃점을 온점으로 잘못 표시하여 중의성이 발생하는데(예: 8.15 (팔점 일오/팔일오)), 소수와 기념일의 구분은 숫자 우 문맥에 나타나는 단어(예: 사태, 전쟁, 공동성명, 선거, 운동, 혁명 등)를 통해 가능하다.

신문데이터에서 각종 기호를 일정한 규칙없이 사용하고 있기 때문에[15] 기호포함 문자열에서 발생하는 중의성 해결은 쉽지 않다. 그러나 정치·경제·사회·스포츠 분야의 전문용어가 포함된 경우가 많으므로 좌우 문맥에 나타나는 연어정보를 이용한다.(예: 3:5로 이기다/꺾다/승리하다/연승하다) 범위수와 부정수의 경우, 수사 크기, 분류사의 종류에 따라 숫자 및 기호의 발음결과가 다양하지만 어느 분야에서나 연어정보없이 사용되어 문맥 정보이용에 한계가 있다.(예: 1950-1960(천구백오십년에서 천구백육십년까지/천구백오육십/천구백쉰에서 천구백예순/천구백오십에서 천구백육십))

(b) '①숫자 전치어 + ②기호포함 숫자열' 구조의 ①이 '그림', '표', '그래프' 등이면 그림, 표, 그래프의 수를 나타낼 수도 있고 순서를 나타낼 수도 있다. (예: 그림 1(그림 하나/그림 일))

(c) '②기호포함 숫자열 + ③문자 단위표현'의 결합구조는 원래 중의성이 없으나 ② + (③) + 일반 의존명사 '만' + 조사 '은/의' 의 결합에서는 중의성이 생길 수 있다. ②가 숫자 '1'일 때, '만'이 일반 의존명사인지 숫자 단위표현인지에 따라 숫자 '1'의 발음 여부도 달라진다.(예: 1만은(일만은) = 1 + 일반 의존명사 '만' + 조사 '은'/ 1만은(만은) = 1 + 단위표현 '만(萬)' + 조사 '은')

(d) '②기호포함 숫자열 + ⑤분류사'의 결합은 분류사 계통에 따라 ㉠고유어 수사 + 고유어 분류사 ㉡고유어 수사 + 한자어 분류사, ㉢한자어 수사 + 한자어 분류사, ㉣고유어 수사 + 외래어 분류사, ㉤한자어 수사 + 외래어 분류사와 같은 구조를 이룰 수 있다. 그런데 ㉡과 ㉢에서 보듯이 한자어 분류사 중 '개', '개과', '개관', '개항', '건', '명', '종', '평', '평형', '표', '항목' 등은 수사의 크기가 일반적으로 100 미만이면 고유어 수사를 선호하고, 100 이상이면 한자어 수사를 선호한다. 또한 고유어 분류사의 경우, ㉠과 같은 구조가 일반적이다.[16]

②의 숫자가 서수인지 기수인지에 따라 의미가 달라지는 분류사는 '구(具), 구(球), '권(券)', '권(券)', '기(基)', '기(期)', '단', '단(段)', '대(臺)', '대', '동(棟)', '동', '반(班)', '번(番)', '장(張)', '점(點)', '조(條)', '척(隻)', '첩(貼)', '첩', '단계(段階)', '단지(단지)', '문항', '세대' 등이 있다. 이러한 구조의 중의성에 의미적 차이를 주는 요소로 분류사 후치어와의 결합관계 및 연어정보가 있다.

⑥분류사 후치어 중 '당', '째', '짜리', '대', '씩'이 붙어서 '권', '편', '장', '번' 등의 분류사에 서수와 기수의 의미 차이를 줄 수 있다. <표 7>은 대상 신문자료에서 조사된 '권', '번', '편' 등에 '째'나 '짜리', '씩'이 붙는 경우 전치한 수사의 발음의 수치이다.[17]

<표 7: 분류사 후치어와의 결합관계>

후치사	짜리			째			씩		
수사 / 분류사	Kca_n, Kca_ni	Cca_b [+u]	total	Kca_n, Kca_ni	Cca_b [+u]	total	Kca_n, Kca_ni	Cca_b [+u]	total
권	3	0	3	0	1	1	2	0	2
번	0	0	0	127	0	127	7	0	7
편	1	0	1	1	0	1	2	0	2
평	10	1	11	0	0	0	0	0	0
장	3	0	3	1	0	1	0	0	0
total	17	1	18	129	1	130	11	0	11

조사된 바와 같이 '권', 번', 편', '평', '장' 등의 분류사에 분류사 후치어로 '짜리', 째', '씩'이 오면 수사의 의미가 90% 이상의 확률로 기수가 됨을 알 수가 있다. 특히 분류사 '권', '장', '번' 등의 숫자 앞에 '제'가 오면 모두 서수의 의미를 가지고, '번'은 좌우 문맥으로 '도

[15] 하나의 기호를 다양한 의미로 사용할 뿐만 아니라 하나의 의미를 다양한 기호로 표현하고 있다.(예: '3-4위전', '3~4위전', '3, 4위전', '3_4위전', '3.4위전', '3·4위전')

[16] 단위수나 십단위 이상의 부정수, 큰 수로는 한자어 수사가 선호되는 경향이 있다. (예: 50그루(오십/쉰 그루), 30-40그루(삼사십 그루), 3459그루(삼천사백오십구/삼천사백쉰아홉 그루))

로(국도, 지방도로)', '시드', '우드', '출구', '버스' 등이 오면 모두 서수의 의미를 가진다.

5. 결론

지금까지 본 연구에서는 현대 한국어에서 아라비안 숫자를 읽는 방식에 대해 알아보았다. 본 연구 결과 아라비안 숫자 읽기는 크게 고유어 방식과 단위를 붙여 읽거나 붙이지 않고 읽는 한자어 방식, 단(單)단위에서의 영어로 읽는 방식으로 분류됨을 알 수 있었다. 그러나 이렇듯 다양하게 읽히는 숫자의 문자화를 단순한 규칙을 설정하여 자동화하기에는 중의성이 높아 본 연구에서는 ① 숫자 전치어, ② 기호를 포함한 숫자열의 표현 형식과 크기, ③ 단위 표현, ④ 숫자 후치어, ⑤ 분류사, ⑥ 분류사 후치어, ⑦ 수사표현 앞뒤 문맥을 포함하여 패턴화 되고 중의성이 없는 아라비안 숫자 읽기 구조를 제시하였다. 또한 규칙을 설정할 수 없는 중의성이 있는 구조는 그 유형을 밝히고 신문의 정치·경제·스포츠 등 각 분야별로 각 문맥 정보가 상이한 의미를 가지기 때문에 이를 목록화하였고 분류사 후치어와의 결합관계를 통해 의미적 차이를 주는 요소를 밝혔다.

그러나 수사의 크기에 따라 고유어 수사, 한자의 수사 및 외래어 수사가 이용되는 범위가 다른데, 이 기준이 명확하지 않고 개별 분류사와 수사구조를 연구하는 학자에 따라 달라져 이러한 교차 수사를 허용하는 분류사에 대한 연구와 함께 각 분류사가 고유어 수사나 한자어 수사와 각각 결합하는 통계수치 정보에 대한 연구가 함께 이루어질 부분이다. 또한 전체 숫자 표현 중 단위를 포함한 한자어로 읽히는 경우가 약 86% 정도인데 중의성 해결이 안되는 숫자 표현의 발음에 이러한 선호도를 적용하였을 때, 결과의 정확도가 어느 정도일지 실험을 통해 밝혀져야 한다.

무엇보다 본 논문에서 언어학적 측면에서 기술된 숫자 읽기 규칙을 실제 신문데이터의 자동처리에 적용하려면, 숫자읽기 규칙이 순차적으로 데이터에 적용되는 알고리즘이 만들어져야하며, 알고리즘의 데이터 적용률과 정확도 분석이 이루어져야 하는데 알고리즘에 관한 연구는 향후 과제로 남긴다.

Acknowledgement

본 논문은 과학기술부의 국가지정연구실사업(과제명: 언어 중심의 지능적 정보처리를 위한 단계적(scalable) 우리말분석기술의 개발(M10203000028-02J0000-01510))의 지원을 받아 이루어졌음을 밝힌다.

참고문헌

[1] 고성철(1990), "국어 수량사구 연구"(석사학위논문), 고려대 대학원

[2] 김상준(1986), "방송과 수의 표현", 『KBS 표준방송언어』, 한국방송공사, 서울.

[3] 김상준(1992), 『방송언어연구:한국어 음성표현의 이론과 실제』, 도서출판 홍원, 서울.

[4] 김병주(2000), "정보인식을 위한 고유명사 및 수사추출"(석사학위논문), 영남대학교 대학원

[5] 김준홍(2000), "도합유사도를 이용한 한국어 추출요약 시스템"(석사학위논문), 한국해양대 대학원

[6] 김영희(1976), "한국어 수량화 구문의 분석", 『언어』,제12호, 한국언어학회, pp. 89-112

[7] 유재원(1999), "자연어 처리를 위한 수사의 하위 범주 분류", 『언어와 언어학』 제24호, pp. 103~110

[8] 유재원(1997), "자연어 처리를 위한 의존명사의 하위 범주 분류", 제9회 한글 및 한국어 정보처리 학술대회 학술발표 논문집, pp. 136~142

[9] 이영직(2000), "방송 뉴스 전사 문장의 수사 및 단위의 발성 방식", 제17회 음성통신 및 신호처리 학술대회, pp.285-288

[10] 이은정(1993), 『최신 표준어·맞춤법 사전』, 백산출판사, 서울.

[11] 채완(1983), "국어 수사 및 수량사구의 유형적 고찰", 『어학연구』 제19권 제1호, pp.19~34

[12] 채완(1990), "국어 어순의 기능적 고찰", 『동대논총』 제20집, pp.103~119

[13] 채완(1998), "의존명사의 사전적 처리" 『새국어생활』, 제8권 제1호, pp.49-63

[14] 황경수(1997), "현대국어의 의존명사 연구", 『청주대 인문과학논집』, 제17집 pp.429-458