

한국코퍼스언어학회 | 한국영어학회
2023년 여름 공동 학술대회 발표논문집

현대 언어학의 현황과 과제: 코퍼스, 빅데이터, 인공지능

**Proceedings of
2023 Summer Joint Conference**
of
the Korean Association for Corpus Linguistics (KACL)
and the Korean Association for the Study
of English Language and Linguistics (KASELL)
on

**State-of-the-Art and Challenges of Modern Linguistics:
Corpus, Big Data, and Artificial Intelligence**

일시: 2023년 6월 3일(토)

장소: 고려대학교 문과대학 서관 2층

주최: 한국코퍼스언어학회 <https://www.corpus.or.kr/>

한국영어학회 <http://www.kasell.or.kr/>

주관: 고려대학교 언어정보연구소

4단계 BK21 딥러닝시대의 전문언어인력양성 사업팀

조직 및 운영위원회

대회장

류병래(충남대), 이석재(연세대)

조직위원회

조직위원장: 송상헌(고려대)

조직위원: 김기택(서울대), 이용훈(충남대), 안성민(KAIST), 성제현(공주대),
노강산(고려대, 조직간사)

프로그램위원회

프로그램위원장: 강아름(충남대)

프로그램위원: 김광섭(한국외대), 김일환(성신여대), 김지은(가톨릭관동대), 안
병길(경상국립대), 안현기(서울대), 오윤정(대만국립정치대학교), 위혜경(단국
대), 윤영은(이화여대), 이문우(한양대), 이재영(서울대), 윤태진(성신여대)

프로그램

09:20-09:30	Opening Remarks (문과대학 서관 202호) 류병래 (충남대학교, 한국코퍼스언어학회장) & 이석재 (연세대학교, 한국영어학회장)		
09:30-10:20	Invited Talk 1 (문과대학 서관 202호) 클라우드 기반 자연어처리의 동향과 전망 Speaker: 김대근 (AIML 전문솔루션즈 아키텍트) Moderator: 송상현		
10:20-10:30	Break		
	Session 1 문과대학 서관 202호 사회: 박슬기 (경희대)	Session 2 문과대학 서관 224B호 사회: 김옥기 (경희대)	Session 3 문과대학 서관 225호 사회: 강아름 (충남대)
10:30-10:55	박윤규, 김종복 (경희대) English all-cleft constructions: A corpus-based approach 토론: 이승아 (이화여대)	손지연 (연세대) 뉴스 인터뷰 대화 내 '네/예'의 기능과 특성 - 인터뷰어 발 화를 중심으로 토론: 이지은 (고려대)	유성현 (광운대) 한국에서의 다국어 사 전 여명기 - 지식영 본 『아학편』(1908) 을 중심으로 토론: 이석재 (연세대)
10:55-11:20	나익주 (전남대) 합성어 '국뽕'의 의 미 해석과 은유 토론: 홍승혜 (고려대)	이지은, 황동진, 송상현 (고려대) 인공지능 언어 모델의 절 연결에 대한 일고 찰 - 이유를 나타내 는 '-어서'와 '- 니까'를 중심으로 - 토론: 이용훈 (충남대)	홍준선 (서울대) On the Diverse Paradigm of Rising Declaratives in English 토론: 강아름 (충남대)

11:20-11:50	Guest Talks (문과대학 서관 202호) LXPER AI 기술 및 영어교육 플랫폼 소개: 생성형 AI 기반의 난이도 구분 지문 생성과 문항 생성 Speaker: 이형중 (LXPER 대표) Moderator: 이석재 (연세대)		
11:50-12:20	연구윤리 교육		
12:20-12:50	Lunch Break		
12:50-13:20	학회별 총회 및 상임이사회		
	Session 1 문과대학 서관 202호 사회: 서혜진 (동국대)	Session 2 문과대학 서관 224B호 사회: 홍승혜 (고려대)	Session 3 문과대학 서관 225호 사회: 박현아 (고려대)
13:20-13:45	김선웅 (광운대학교) On Repair by Subject Parasitic Gaps 토론: 전해원 (한국해양대)	Liu Cuilin, 박호민, 장세은 (한국해양대) A Corpus-based Multi-lingual Comparison of AI-based Machine Translation: English, Korean, and Chinese 토론: 송상현 (고려대)	성민창 (경인교대), 김은형, 김기택 (서울대) 초등영어 학습자 어휘 수준에 적합한 아동문학 작품 탐색 연구 토론: 안성민 (KAIST)
13:45-14:10	유은영, 이용훈, 강아름 (충남대) 적대적 사례 생성을 위한 문장 내 단어 교체 알고리즘 토론: 이규민 (고려대)	Zhang Guandong (한국해양대) A corpus-based analysis of lexical bundles in maritime legal genres 토론: 류병래 (충남대)	정윤희 (고려대) 한국어능력시험 II 간접 말하기 문항과 간접 행동 문항 오답지의 화용적 오류 유인 요소 분석 토론: 김기택 (서울대)
14:10-15:00	Invited Talk 2 (문과대학 서관 202호) Corpus Linguistics and Digital Humanities Speaker: 장세은 (한국해양대) Moderator: 류병래 (충남대)		

15:00–15:10	Break		
	Session 1 문과대학 서관 202호 사회: 이지은 (고려대)	Session 2 문과대학 서관 224B호 사회: 김정수 (경희대)	Session 3 문과대학 서관 225호 사회: 강아름 (충남대)
15:10–15:35	윤인경, 유석훈 (고려대) Korean Middle School Third-Year English Textbook Reading Passage Difficulty Analysis: Focusing on Syntactic Complexity and Readability 토론: 박현아 (고려대)	전해원 (한국해양대) Clausal Ellipsis and Non-Simultaneous Transfer 토론: 김정수 (경희대)	이용훈, 이예빈 (충남대) Examining the Transformer Models with the K-CoLA Dataset 토론: 신운섭 (고려대)
15:35–16:00	류병래 (충남대) 한국어 반접어 (anti-clitic) 와 통사적 파생: ‘것’, ‘놈’, ‘자’ 를 중심으로 토론: 장세은 (한국해양대)	최아녕 (연세대) A study on sentence stress and isochrony produced by Korean EFL Learners 토론: 김지은 (가톨릭관동대)	김원빈 (연세대) Meeting between artificial intelligence and corpus linguistics: Collocation analysis for method validation of AI 토론: 이용훈 (충남대)
16:00–16:10	Break		
16:10–17:00	Invited Talk 3 (문과대학 서관 202호) 챗GPT, 언어학 시험을 보다: 협동연구 보고 Speaker: 송상현 (고려대) Moderator: 강아름 (충남대)		
17:00–17:10	Closing Remarks (문과대학 서관 202호) 유석훈 (고려대학교 언어정보연구소장, 4단계 BK21 딥러닝시대의 전문 언어인력양성 사업팀장)		

개회사

한국코퍼스언어학회와 한국영어학회의 2023 여름 공동학술대회를 오늘 이곳 유서깊은 고려대학교에서 개최하게 되어 매우 기쁘게 생각합니다. 이번 학술대회의 주제는 “현대 언어학의 현황과 과제: 코퍼스, 빅데이터, 인공지능”입니다. 최근 챗GPT(ChatGPT)와 바드(Bard)로 대표되는 대규모 언어 모델(Large Language Model, LLM) 기반 생성형 AI에 대한 관심이 뜨겁습니다. 자연언어처리 기술과 대용량 언어자료의 중요성이 부각되는 요즘 이 학술대회는 관련 주제의 최신 연구 결과를 공유하고 최신 학술정보를 주고 받는 소중한 소통의 장이 될 것으로 확신하며 기대 또한 매우 큽니다.

오늘 열리는 학술대회에서는 초청강연 3편, 객원 강연 1편을 포함하여 총 22편의 논문이 발표됩니다. 먼저 초청 연사이신 김대근, 장세은, 송상헌 선생님께 깊은 감사의 말씀을 올립니다. 또한 일반 논문 발표자, 토론자, 사회자 그리고 참여한 모든 분들께 감사를 드립니다.

이번 공동 학술대회를 같이 준비하신 한국영어학회 이석재 회장님과 양 학회 임원진 여러분의 노고에 뜨거운 찬사와 깊은 감사의 마음을 전합니다. 특히 양 학회 총무이사이신 김기태 교수님과 이용훈 교수님께 특별한 감사의 마음을 전합니다.

또한 이번 학술대회를 위해 물심양면으로 도움을 주신 고려대학교 관계자 여러분, 특히 주관을 해주신 고려대학교 언어정보연구소와 4단계 BK21 딥러닝시대의 전문 언어인력양성 사업팀장 유석훈 교수님, 조직위원장 송상헌 교수님과 팀원 여러분께 감사의 말씀을 올립니다. 그리고 훌륭한 프로그램을 만들어 알찬 학술대회가 되도록 최선을 다해 주신 강아름 위원장님께 특별히 깊은 감사를 드립니다. 또한 후원을 해주신 (주)메트릭스 나윤정 대표님께도 감사의 인사를 올립니다.

이번 학술대회에 참석하신 모든 분들이 오늘 열리는 지적 향연을 즐기시기를 바라고 참석하신 모든 분들에게 학문적으로 큰 성취가 있으시길 간절히 기원드립니다.

2023년 6월
한국코퍼스언어학회 회장

류병래

모시는 글

안녕하십니까, 한국코퍼스언어학회와 한국영어학회의 2023년 여름 공동 학술대회에 여러분을 모시게 되어서 기쁜 마음 그지없습니다.

작금 언어학의 연구 방향과 그 방법은 큰 궤도 수정이 불가피한 상황으로 여겨질 수밖에 없는 것 같습니다. 이제 인간 본연의 고유 능력에 속한다고 여겨지던 언어를 인공지능이 학습하고 생성하는 시대에 와 있습니다. 인간 마냥 잘 못 배우면 사실과 다른 이야기도 하고 아직 실수도 합니다만, (올바른 학습을 충분히 거친다면) 자율적으로 생각하고 판단하는 통합적 의미 생산 주체로서의 기계가 탄생한 것입니다. 언어학도들에게는 이제 언어의 분석을 통해서 인간의 사고와 표출의 표상을 연구하는 것만큼 언어 그 자체와 사회, 문화, 교육, 예술, 경제, 과학 등등 언어가 관여된 세상 지식을 인공지능이 학습할 수 있도록 언어를 조련하는 역할의 중요성도 커졌다고 봅니다. 언어 연구에 있어 분석도 중요하지만 이제 분석된 자료를 언어 내외적으로 종합하는 능력이 더 요구될 것으로 조심스럽게 예견해 봅니다. 한 언어 내적으로는 어휘, 의미, 통사, 음운의 종합적 연구가 더욱 필요해 보이고, 언어 간에 있어서도 종합 대조 연구가 더욱 절실하고, 무엇보다도 사회의 각 분야에서의 언어 사용에 대해 더욱 촘촘한 연구와 데이터 확보가 중요한 시점인 것 같습니다.

한국코퍼스언어학회는 이처럼 언어 연구에 있어 언어 빅데이터의 중요성을 숙지하고 이를 언어의 계량적 연구에 직접적으로 관여시키고 있고, 한국영어학회는 여러 언어 중 영어에 초점을 맞추어 언어의 보편적 특징 위에 영어 자체의 언어적 특징과 영어의 교육적 사용 및 응용에 연구력을 집중시키는 학술 단체로서 이미 인공지능 시대의 학술적, 교육 사회적 요구 사항을 적극적으로 반영하는 언어 연구 풍토를 가꾸어 가고 있습니다. 이에 걸맞게 두 학회는 이번 공동 학술대회의 주제를 “현대 언어학의 현황과 과제: 코퍼스, 빅데이터, 인공지능”으로 설정하고, 관련 분야의 최근 주 관심사에 대한 발전을 함께 공부하고 심도 있는 토의를 진행하려 합니다. 언어학 연구의 격변의 한가운데에서 많은 연구자분의 적극적인 참여를 기대하며, 양 학회의 뜻깊은 이번 공동 학술대회의 의의를 되새기셨으면 하는 마음 간절합니다.

이 자리를 빌려 본 학술대회 개최에 많은 도움을 주신 고려대학교 언어정보연구소 4단계 BK21 딥러닝 시대의 전문언어인력양성 사업팀장님이신 유석훈 교수님과 류병래 한국코퍼스언어학회장님, 그리고 준비 실무 책임을 맡아주신 송상헌 공동학

술대회 조직위원장님과 양 학회의 강아름, 이용훈, 김기택, 안성민, 성제현 교수님과
노강산 간사님께 감사의 마음을 전합니다.

2023년 6월
한국영어학회 회장

이석재

환영사

본 학술대회에 참석하신 모든 분들께,

고려대학교에서 한국코퍼스언어학회-한국영어학회 공동학술대회를 개최하게 된 것은 큰 기쁨이라고 하지 않을 수 없습니다. 이번 학술대회는 지난 몇 달 동안 학술대회 조직 위원회와 연구자분들의 많은 노력과 준비 끝에 열리게 되었습니다.

특히 이번 인사 말씀을 통해 대표 발표자인 김대근 선생님, 이형종 선생님, 장세은 교수님, 그리고 송상헌 교수님의 열정과 지식에 깊은 감사를 드리고자 합니다. 그동안 많은 연구자분들께서 준비해 오신 연구 발표에 대해서도 큰 박수를 보내드립니다. 이번 공동학술대회를 통해 우리는 서로의 지식을 나누고 혁신적인 아이디어를 공유하며 성장할 수 있었습니다. 코로나 사태 이후 저희가 오늘 대면으로 함께 한 학술대회 현장은 아름다운 추억으로 간직될 것 같습니다. 이곳에서 얻은 친목과 교류는 우리 모두에게 큰 보람과 영감을 주었습니다.

이번 학술대회의 성공적인 개최는 많은 분들의 노고 덕분에 가능했습니다. 주관 기관, 후원자, 조직위원회, 발표자, 그리고 참석자 여러분들께 감사드립니다. 우리의 협력과 열정이 이룬 성과입니다. 여러분들께서 새로운 아이디어와 지식을 얻고, 서로를 영감과 동기부여의 근원으로 삼아 꾸준한 성장을 이루시기를 바랍니다. 이번 학술대회에서 만들어진 네트워크와 연결은 앞으로의 연구와 학문 발전에 큰 도움이 될 것입니다.

함께 참여하고 지원해 주신 모든 분들께 다시 한번 감사의 말씀을 전하며, 앞으로의 학술적인 도전과 협력을 기대합니다.

감사합니다.

2023년 6월

고려대학교 언어정보연구소장

4단계 BK21 딥러닝시대의 전문언어인력양성 사업팀장

유석훈

차례

조직 및 운영위원회	iii
프로그램	v
개회사	ix
모시는 글	xi
환영사	xiii

Invited Talk 1

클라우드 기반 자연어 처리의 동향과 전망	1
김대근 (AIML 전문솔루션즈 아키텍트)	

Session A

English all-clef constructions: A corpus-based approach	20
박윤규, 김종복 (경희대학교)	
합성어 '국뽕'의 의미 해석과 은유	38
나익주 (전남대학교)	

Session B

뉴스 인터뷰 대화 내 '네/예'의 기능과 특성 - 인터뷰어 발화를 중심으로	42
손지연 (연세대학교)	
인공지능 언어 모델의 '절 연결 능력'에 대한 고찰 -이유를 나타내는 '-어서'와 '-니까'를 중심으	

로-	50
이지은, 황동진, 송상현 (고려대학교)	

Session C

아학편(兒學編) <지석영본>(1908)의 구성과 전사·번역 양상	61
유성현 (광운대학교)	
On the Diverse Paradigm of Rising Declaratives in English	70
홍준선 (서울대학교)	

Guest Talk

생성형 AI 기반의 난이도 구분 지문 및 문항 생성	82
이형중 (LXPER 대표)	

Session A

On Repair of the Subject Condition by Parasitic Gaps	98
김선웅 (광운대학교)	
Homonyms-based word replacement algorithm for generating adversarial examples	
유은영, 이용훈, 강아름 (충남대학교)	109

Session B

A corpus-based Multi-lingual Comparison of AI-based Machine Translation: English, Korean, and Chinese	118
Liu Cuilin, 박호민, 장세은 (한국해양대학교)	

A corpus-based analysis of lexical bundles in maritime legal genres	129
Zhang Guandong (한국해양대학교)	

Session C

초등영어 학습자 어휘 수준에 적합한 아동문학 작품 탐색 연구	139
성민창 (경인경기교육대학교), 김은형, 김기택 (서울대학교)	

한국어능력시험 II 오답지의 화용적 오류 유인 요소 분석 - 간접 말하기 문항과 간접 행동 문항을 대상으로	159
정윤희 (고려대학교)	

Invited Talk 2

Invited Talk 2: Corpus Linguistics and Digital Humanities	176
장세은 (한국해양대학교)	

Session A

Korean Middle School Third-Year English Textbook Reading Passage Difficulty Analysis: Focusing on Syntactic Complexity and Readability	204
윤인경, 유석훈 (고려대학교)	

한국어 반접어(anti-clitic)와 통사적 파생: '것', '놈', '자'를 중심으로	
류병래 (충남대학교)	214

Session B

Clausal Ellipsis and Non-simultaneous Transfer	233
전혜원 (한국해양대학교)	

A Study on Sentence Stress and Isochrony Produced by Korean EFL Learners	243
최아녕 (연세대학교)	

Session C

Examining the Transformer Models with the K-CoLA Dataset	254
이용훈, 이예빈 (충남대학교)	

Meeting between artificial intelligence and corpus linguistics: Collocation analysis for method validation of AI	264
김원빈 (연세대학교)	

Invited Talk 3

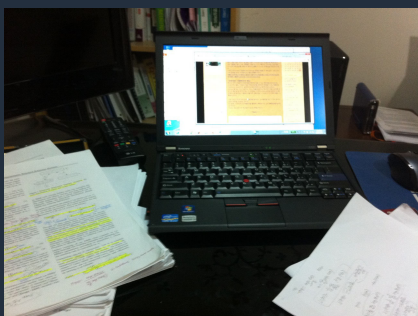
Invited Talk 3: 챗GPT, 언어학 시험을 보다: 협동연구 보고	273
송상현 (고려대학교)	

클라우드 기반 자연어 처리의 동향과 전망

김대근

APJ Senior AIML Specialist Solutions Architect
Amazon Web Services

과거: **Entry barrier** was high



- 수학 & 통계 필수
- No open source
→ 논문 직접 구현

(3) Suppose Given $p(\theta | X, y) = \frac{1}{Z} N(w; \mu, \Sigma)$ (defined by 3.10)

$p(w) = N(w; \mu, \Sigma)$ (by 2.48), we have:

$$p(w | \theta) \propto \frac{1}{Z} N(w; \mu, \Sigma) \prod_{i=1}^n N(w; \mu_i, \Sigma_i)$$

$$= \frac{1}{Z} N(w; \mu, \Sigma) \prod_{i=1}^n N(w; \mu_i, \Sigma_i)$$

(where μ is a NMM design matrix)

$$= C \exp \left(-\frac{1}{2} (w - \mu)^T \Sigma^{-1} (w - \mu) \right) \exp \left(-\frac{1}{2} (w - \mu_i)^T \Sigma_i^{-1} (w - \mu_i) \right)$$

$$\propto \exp \left(-\frac{1}{2} (w - \mu)^T \Sigma^{-1} (w - \mu) \right) \exp \left(-\frac{1}{2} (w - \mu_i)^T \Sigma_i^{-1} (w - \mu_i) \right)$$

$$= \exp \left(-\frac{1}{2} \left[(w - \mu)^T \Sigma^{-1} (w - \mu) + (w - \mu_i)^T \Sigma_i^{-1} (w - \mu_i) \right] \right)$$

$$= \exp \left(-\frac{1}{2} \left[w^T \Sigma^{-1} w - 2w^T \Sigma^{-1} \mu + \mu^T \Sigma^{-1} \mu + w^T \Sigma_i^{-1} w - 2w^T \Sigma_i^{-1} \mu_i + \mu_i^T \Sigma_i^{-1} \mu_i \right] \right)$$

$$= \exp \left(-\frac{1}{2} \left[w^T (\Sigma^{-1} + \Sigma_i^{-1}) w - 2w^T (\Sigma^{-1} \mu + \Sigma_i^{-1} \mu_i) + \mu^T \Sigma^{-1} \mu + \mu_i^T \Sigma_i^{-1} \mu_i \right] \right)$$

$$= \exp \left(-\frac{1}{2} \left[w^T (\Sigma^{-1} + \Sigma_i^{-1}) w - 2w^T (\Sigma^{-1} \mu + \Sigma_i^{-1} \mu_i) + \mu^T \Sigma^{-1} \mu + \mu_i^T \Sigma_i^{-1} \mu_i \right] \right)$$

(Note that $\Sigma^{-1} + \Sigma_i^{-1}$ is a valid covariance matrix)

$$= \exp \left(-\frac{1}{2} \left[w^T (\Sigma^{-1} + \Sigma_i^{-1}) w - 2w^T (\Sigma^{-1} \mu + \Sigma_i^{-1} \mu_i) + \mu^T \Sigma^{-1} \mu + \mu_i^T \Sigma_i^{-1} \mu_i \right] \right)$$

$$= \exp \left(-\frac{1}{2} \left[w^T (\Sigma^{-1} + \Sigma_i^{-1}) w - 2w^T (\Sigma^{-1} \mu + \Sigma_i^{-1} \mu_i) + \mu^T \Sigma^{-1} \mu + \mu_i^T \Sigma_i^{-1} \mu_i \right] \right)$$

(Since $\mu = \Sigma (\Sigma^{-1} \mu + \Sigma_i^{-1} \mu_i)$ and $\Sigma = \Sigma^{-1} \Sigma$ by (2.11))

$$= \exp \left(-\frac{1}{2} \left[w^T (\Sigma^{-1} + \Sigma_i^{-1}) w - 2w^T (\Sigma^{-1} \mu + \Sigma_i^{-1} \mu_i) + \mu^T \Sigma^{-1} \mu + \mu_i^T \Sigma_i^{-1} \mu_i \right] \right)$$

$$= \exp \left(-\frac{1}{2} \left[w^T (\Sigma^{-1} + \Sigma_i^{-1}) w - 2w^T (\Sigma^{-1} \mu + \Sigma_i^{-1} \mu_i) + \mu^T \Sigma^{-1} \mu + \mu_i^T \Sigma_i^{-1} \mu_i \right] \right)$$

Therefore, the result (3.10) is verified. \square

```
void Filtering::MotionEst_DiamondSearch(unsigned char *curimg, unsigned char *refimg)
{
    register int i, j, k;
    int w, gy;
    int xlast, ylast;
    int bi, bj;
    int refBLwer, refBLwer;
    int w = m_width;
    int h = m_height;
    int sum_gy, sum_gx;
    int L = (int) floor(log10(p-1) / log10(2));
    int posin;
    unsigned char SOSFlag, cornerFlag;

    unsigned char *py = new unsigned char[m_height * m_width];
    unsigned char *px = new unsigned char[m_height * m_width];
    unsigned char *refblk = new unsigned char[m_height * m_width];
    unsigned char *curblk = new unsigned char[m_height * m_width];
    float *costs = new float[m_height * m_width];
    for (k = 0; k < 9; k++) costs[k] = 10000.0f;
    memset(mvs, 0, m_width * m_height * sizeof(MotionVector));

    for (i = 1, bi = 0; i <= h - MB_SIZE; i += MB_SIZE, bi++)
    {
        for (j = 1, bj = 0; j <= w - MB_SIZE; j += MB_SIZE, bj++)
        {
            y = i * x = j;

            memcpy(curblk, curimg, w, h, j-1, i-1, MB_SIZE, MB_SIZE);
            gradient(gx, gy, &sum_gx, &sum_gy, curblk, MB_SIZE, MB_SIZE);
            if (sum_gx < GRAD_THRES || sum_gy < GRAD_THRES)
            {
                mvs[(bi * m_width + bj) * sizeof(MotionVector)] = 0;
                continue;
            }

            memcpy(refblk, refimg, w, h, j-1, i-1, MB_SIZE, MB_SIZE);
            costs[0] = costFunction(refblk, curblk, MB_SIZE);

            for (k = 0; k < 9; k++)
            {
                refBLwer = y + LDP[k][1];
                refBLwer = x + LDP[k][0];
                if (refBLwer < 1 || refBLwer > h || refBLwer < 1 || refBLwer > w)
                    continue;

                if (k == 4)
                {
                    memcpy(refblk, refimg, w, h, refBLwer-1, refBLwer-1, MB_SIZE, MB_SIZE);
                    costs[k] = costFunction(refblk, curblk, MB_SIZE);
                }
            }
        }
    }
}
```

Source: Presenter(Daekeun)'s own materials

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

현재: **Entry barrier** is low



Chip Huyen · Following

Real-time ML @ Claypot AI | ML Sys @ Stanford | Hiring strong streaming engin...
6d · 🌐

Now is the time to get into AI

In 2016, Google launched its AI residency program with around 20 people in their first cohort. They all went on to do amazing research, making ChatGPT and Claude possible, or launching startups like [Character.AI](#), Adept, Harvey.

A few weeks ago, I caught up with [David Ha](#), who was in the first AI residency cohort and is now head of research at Stability AI. I asked him: "How did Google do it? What made their first cohort so successful?"

"Timing," David told me.

2016 was the year when many breakthroughs in AI research were just shaping up: attention mechanism, transfer learning, architecture search, distributed training. Nobody was yet an expert. It was possible for smart and motivated newcomers to quickly catch up and establish themselves.

I see the same thing happening with generative AI.

Everything is new. Nobody is yet an expert. There's not yet a mountain of established literature to waddle through. There are so many possibilities, and the entry barrier is low.

Now is the time to jump in.

[Inspired by a tweet by [Chris/Albon](#)]

👍👍👍 You and 3,625 others

93 comments · 113 reposts

Update Logs

- 2023.04.20
 - KoAlpaca v1.1 데이터셋 공개
 - Polyglot-ko 12.8B 기반 KoAlpaca-Polyglot-12.8B-v1.1b fp16 모델 및 8bit 모델 공개
 - ChatKoAlpaca Web UI(v1.1b) 공개
- 2023.03.26: LLAMA 30B 기반 KoAlpaca 모델을 공개합니다. (LoRA로 학습)
 - LLAMA 30B 학습은 GIST Sundong Kim 교수님의 A100 지원으로 학습되었습니다. 감사합니다 🙏
- 2023.03.24: LLAMA 13B 기반 KoAlpaca 모델을 공개합니다. (LoRA로 학습)
- 2023.03.23: LLAMA 65B 기반 KoAlpaca 모델을 공개합니다. (LoRA로 학습)
- 2023.03.22: 카카오톡에 포팅된 KoAlpaca 봇이 추가되었습니다.

KoAlpaca: Korean Alpaca Model based on Stanford Alpaca (feat. LLAMA and Polyglot-ko)

Stanford Alpaca 모델을 학습한 방식과 동일한 방식으로 학습을 진행한, 한국어로 이해하는 Alpaca 모델입니다.

바로 써보기: 채팅형 웹 페이지(v1.1b-chat) 혹은 카카오톡 봇, Telegram Bot으로 만나보세요!

[ChatKoAlpaca: 채팅형 KoAlpaca \(Polyglot-ko-12.8b-v1.1b\)](#)

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

잘못된 믿음

백두산의 높이는 _____



2,744m

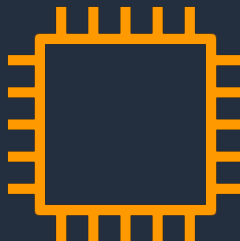
사람처럼 자연어를 이해 understanding하는 것일까요?

2023

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

실제

백두산의 높이는 _____



후보 확률

2,744m	95%
270	3%
한라산	1%
3,000	0.5%
....	...

수많은 후보로부터 가장 높은 확률값을
가지는 후보를 후처리해서 자연어 형태로 가공해 반환



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

단순 규칙의 한계: 번역 예시

You had a haircut

너는 가졌다 머리를 자르는
것을

너 머리 잘랐구나.



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

신경망Neural Network

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

신경망 기반 기계번역

인코더Encoder

You → had → a → haircut → [END]

문장 압축

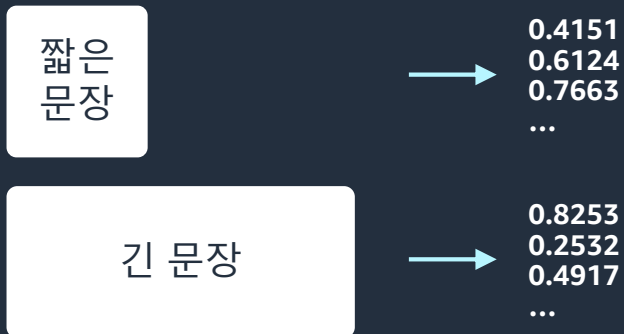
디코더Decoder

너 → 머리 → 잘랐구나 → [END]

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

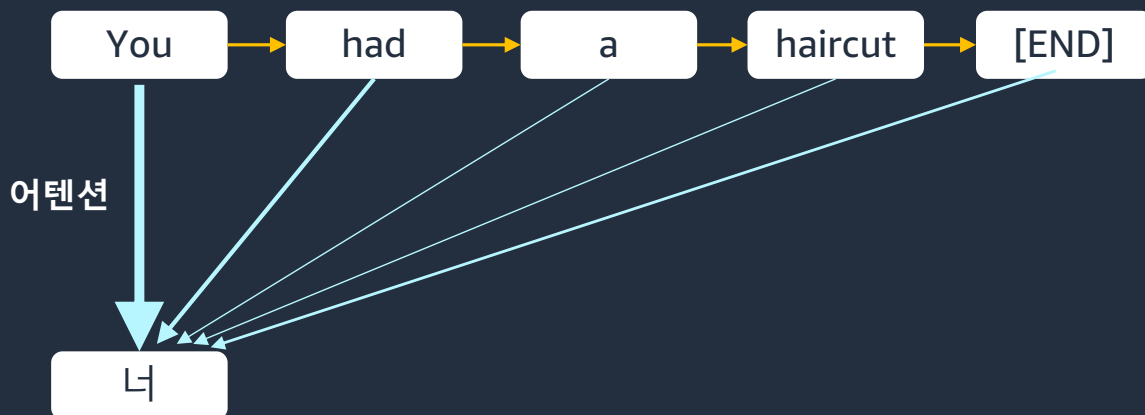
한계점

- 문장 길이가 매우 길다면 압축이 잘 될까요?
- 중요한 단어/구문을 알 수 있나요?
- 신조어가 등장한다면?



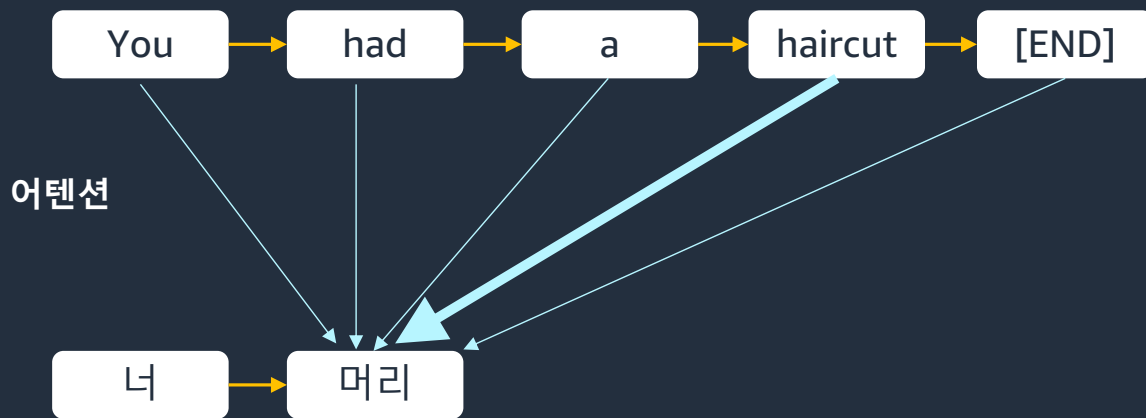
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

혁신의 태동: 어텐션 Attention



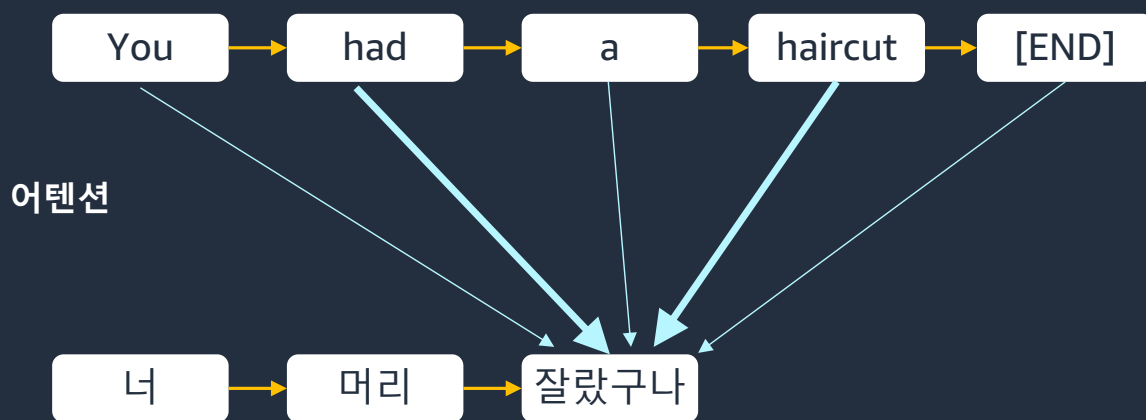
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

혁신의 태동: 어텐션 Attention



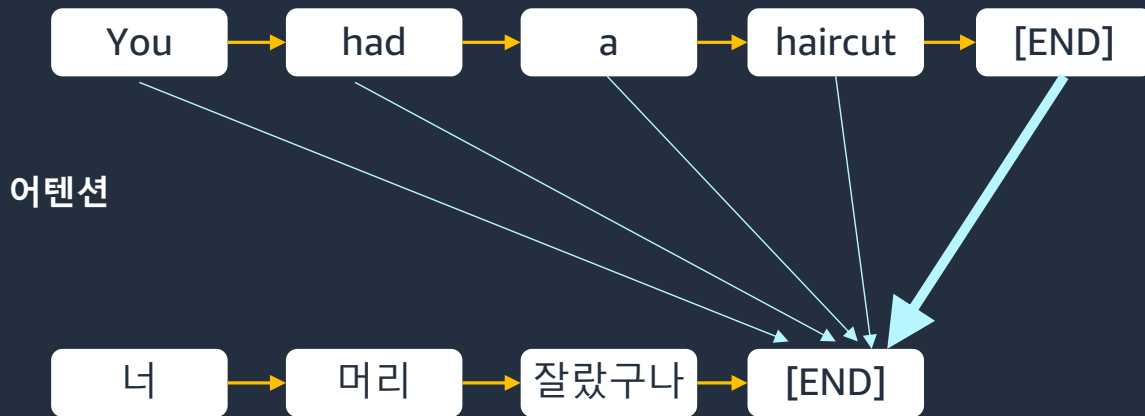
© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

혁신의 태동: 어텐션 Attention



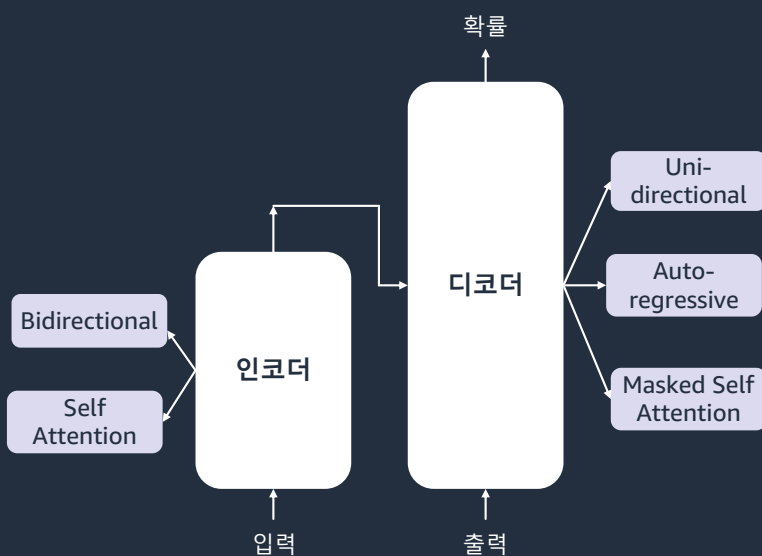
© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

혁신의 태동: 어텐션 Attention



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

생성 AI의 대두: 트랜스포머 Transformers

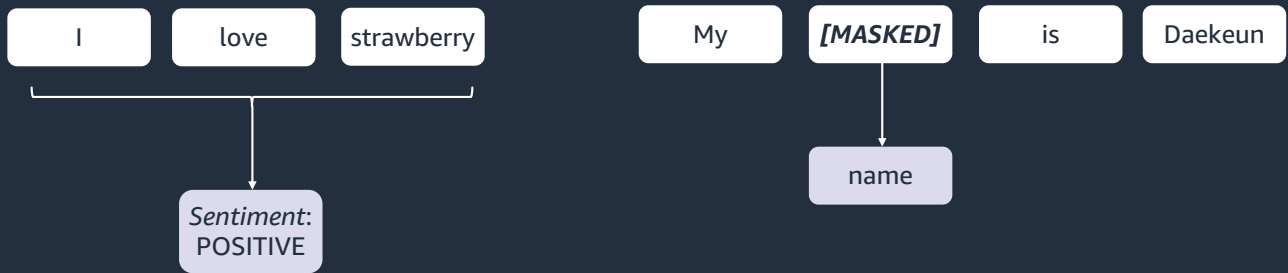


- 순차 데이터간의 관계를 **병렬** 연산으로 처리
- 인코더: 입력 데이터의 **정보를 압축**하는 '**피처 벡터**' 구축
- 디코더: 인코더가 압축한 피처 벡터로 **문장 생성**

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

인코더 모델

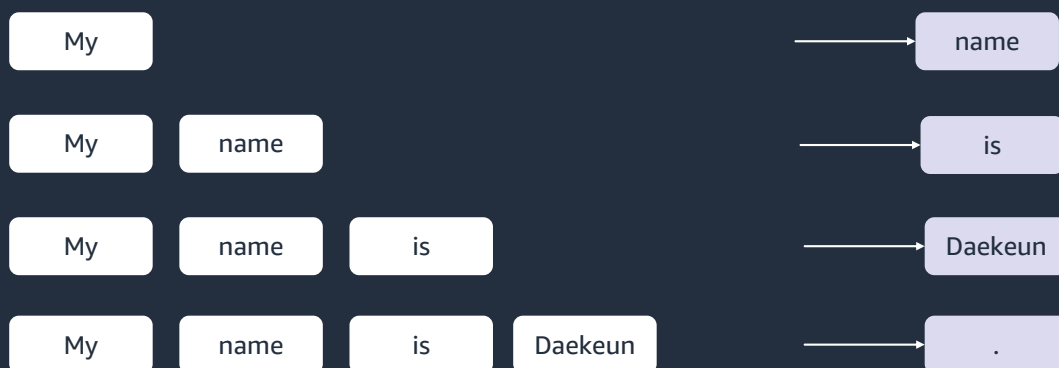
- 언어 이해 모델
- 입력 문장의 모든 단어에 액세스 가능
- 문장 분류, 개체명 인식, 마스크 모델링, 질문/답변 등 입력에 대한 이해가 필요한 작업에 적합
- 예: BERT, ALBERT, DistilBERT, RoBERTa 등



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

디코더 모델

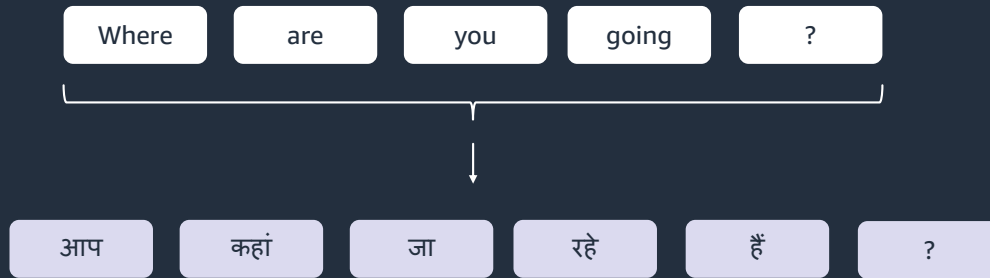
- 문장 생성 모델
- 현재 처리 중인 단어 앞의 단어들만 액세스 가능
- 이전 문맥이 주어진 상태에서 다음 단어를 예측하는 작업에 적합
- 예: GPT, OPT, BLOOM, LLaMA 등



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

인코더-디코더 모델

- 임의의 텍스트 일부분을 마스크 단어로 대체하고 이 마스크가 대체할 텍스트 예측
- 문장 요약, 번역 등과 같은 작업에 적합
- 예: BART, T5, UL2, Flan-T5 등



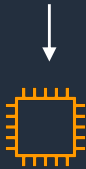
18

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

디코더 모델의 난제 (프롬프트 엔지니어링)

GPT-3

6살짜리 아이에게
달 착륙을 설명하세요.



양자역학은...
(할루시네이션HALLUCINATION)

GPT-3 w/ Prompt

6살짜리 아이에게
달 착륙을 설명하세요.

다음은 배경 지식에 대한 예시입니다.

최초로 달에 발을 내딛은 사람은 닐
암스트롱입니다.

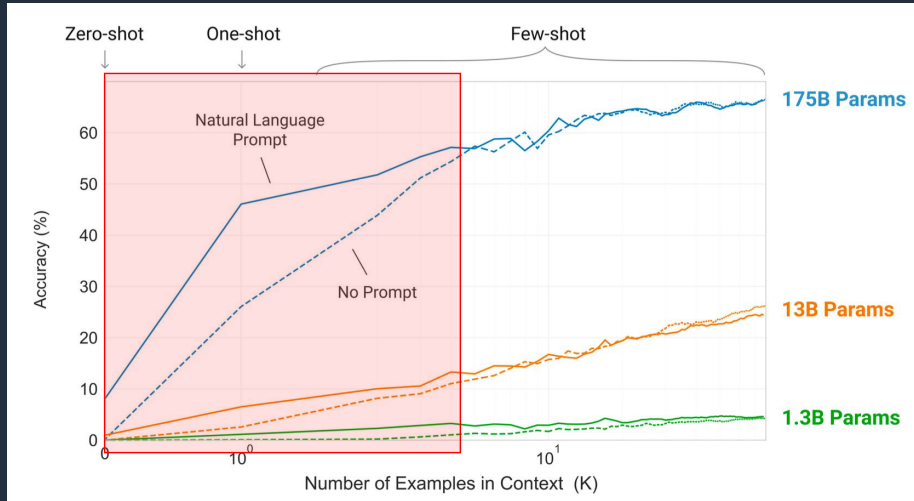
1969년 7월 아폴로 11호가....



아빠가 태어나기 훨씬
전에 우주 비행사가...

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

프롬프트 엔지니어링의 의존성 증가



Source: "Language Models are Few-Shot Learners", 22 July 2020, OpenAI

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

챗GPT: 게임 체인저

Step 1: SFT 모델

6살짜리 아이에게 달 착륙을 설명하세요.

👤 우주선이 달에...

⚙️ 지도 미세 조정
(SFT^{Supervised Fine Tuning})

Step 2. 보상 모델

6살짜리 아이에게 달 착륙을 설명하세요.

👤 달 착륙은 👤 암스트롱...
A ... B

👤 인생은... 👤 빅뱅...
C D

⚙️ A > B > D > C
보상 모델 훈련

Step 3. 강화 학습

세종대왕에 대한 이야기를 해주세요.

SFT 모델

세종대왕은...

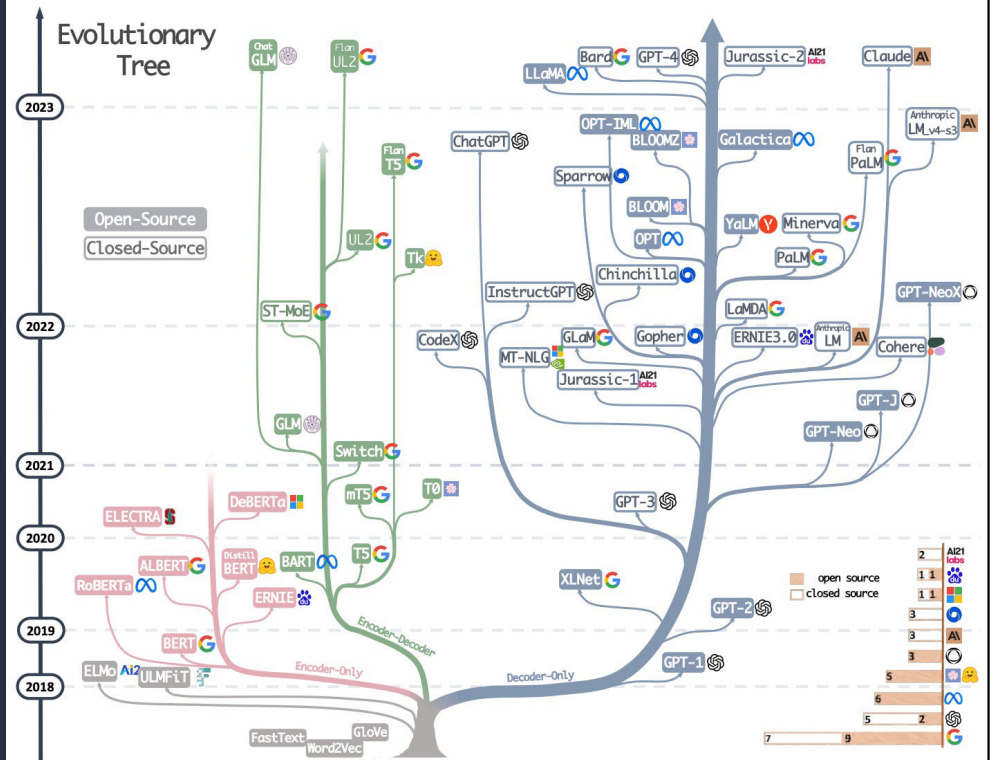
보상 모델

정책 업데이트

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

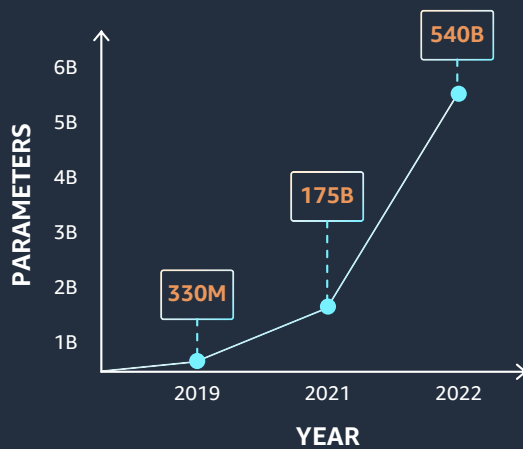
초거대 언어 모델 Large Language Models; LLM의 진화 트리

Source: JingFeng et al. "Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond", arXiv:2304.13712 (2023)



초거대 모델 크기의 급격한 증가

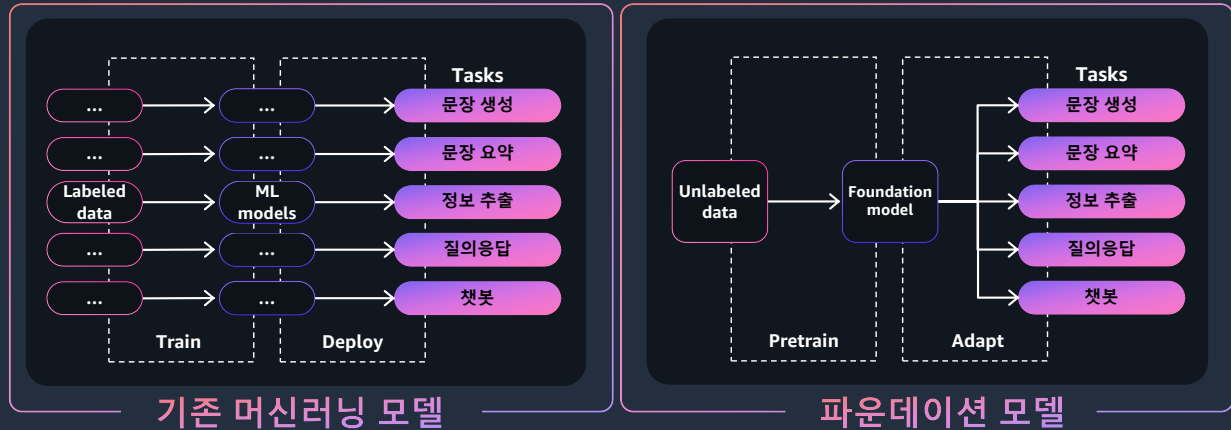
파라미터 개수의 급격한 증가 → 3년 간 2700배 증가



모델	Parameters	year
BERT	340 M	2019
GPT-2	1.5 B	2019
GPT-3	175 B	2020
GPT-NeoX	20 B	2021
Dall-e 2	3.5 B	2021
Stable diffusion	890 M	2022
PaLM	540 B	2022
Bloom	176 B	2022
Jurassic-2	178 B	2023

2021~2023: 파운데이션 Foundation 모델의 대중화

- 2021년: 스탠포드대 HAI 센터에서 파운데이션 모델이란 용어 대중화 (출처: 위키백과)
- 2022년: 파운데이션 모델 (Stable Diffusion, 챗GPT) 의 대두
- 2023년: 최소 파라미터만으로 파인튜닝을 수행하는 다양한 기법 대두



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

파운데이션 모델의 범주

Large Language Models (LLMs)

Foundation Models

LLM:
GPT
Titan
Llama
LaMDA
PaLM

Vision:
SAM
DINO
Diffusion

Sound/Voice:
SpeechT5,
Wav2Vec-U

Multi-Modal

LM+Vision:
CLIP
LLAVA
Midjourney
Stable
Diffusion

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

통념

- 기존 1년이 걸리던 작업을 1주만에 할 수 있다.
- 챗GPT가 한국어가 잘 되니 다른 모델도 한국어가 잘 된다.
- 이제 프롬프트 엔지니어링같은 가내수공업이 필요 없다.
- 파운데이션 모델로 대부분의 태스크를 풀 수 있다.
- 파운데이션 모델로 모든 답변이 가능하다.
- 오픈소스가 넘쳐나니 머신러닝 시스템을 쉽게 구축할 수 있다.

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

초거대 모델 훈련의 난제

* 실제 비용이 아닌 추정치입니다.

\$2-12 million per training for large model

- 막대한 컴퓨팅 리소스 → 인프라 구축 및 훈련에 필요한 비용
- GPT-3의 단일 훈련 세션 비용 : 약 1.4 million 달러
- 대규모 언어 모델 훈련 비용 : 200만~1200만 달러.
- 1월 평균 1,300만 명의 ChatGPT 순 방문자에 대한 칩 요구 사항은 Nvidia A100 GPU 30,000개 이상
- 초기 투자 비용은 약 8억 달러, 일일 전기 요금은 약 50,000달러

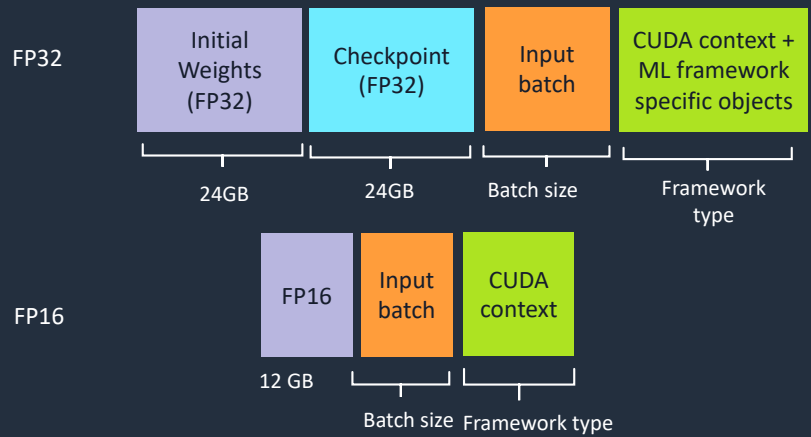
<https://www.techgoing.com/how-much-does-chatgpt-cost-2-12-million-per-training-for-large-models/>

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

초거대 모델 배포의 난제 (1)

- 1장의 GPU로 처리가 어려움
 - GPT-J 6B 모델도 겨우 가능
 - GPT-3 (175B) 규모 모델은?
- 느린 추론 속도

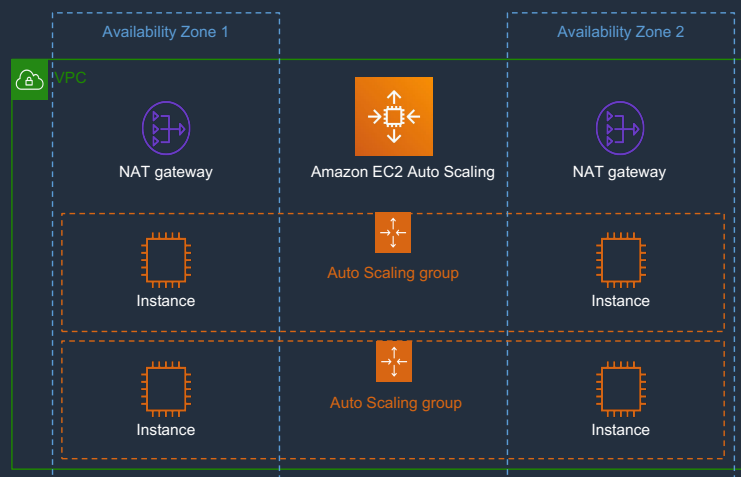
Model serving (GPT-J 6B)



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

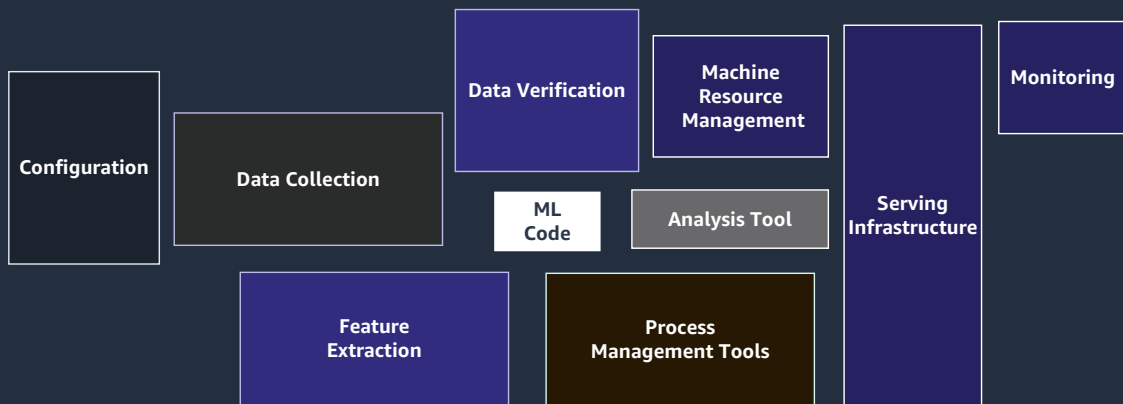
초거대 모델 배포의 난제 (2)

- 인프라 설정
- 고가용성
- 프레임워크 버전 관리?
- 트래픽이 몰린다면?
- A/B 테스트
- 보안을 고려하려면?
- 비용을 아끼고 싶은데?
-



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

머신러닝 시스템 구축의 난제



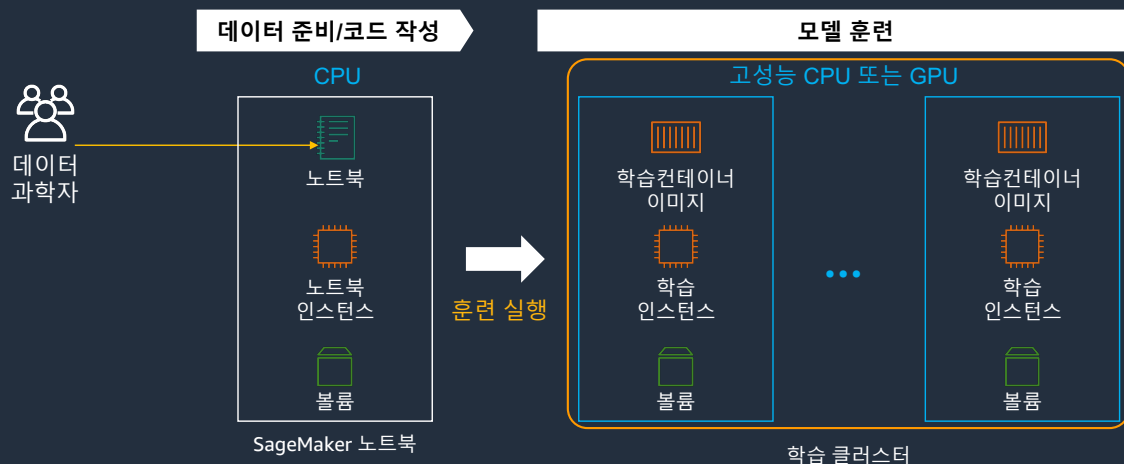
"Only a small fraction of real-world ML systems is composed of the ML code. The required surrounding infrastructure is vast and complex."

source: Hidden Technical Debt in Machine Learning Systems [D. Sculley, & al.] – 2015
<https://proceedings.neurips.cc/paper/2015/file/86df7dcfd896fc2674f757a2463eba-Paper.pdf>

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

클라우드의 이점 (모델 훈련)

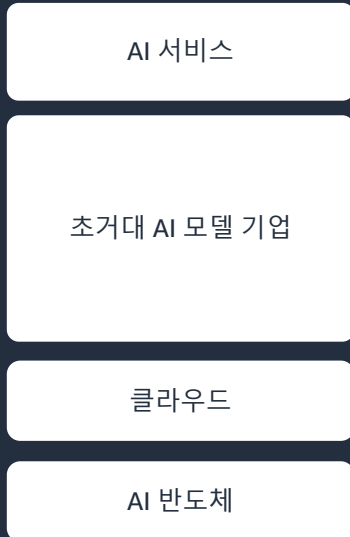
다수의 인스턴스로 확장하여 모델 훈련 가능 (단, 훈련 코드는 분산 훈련 라이브러리 적용 필요)



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

생성 AI의 생태계

대분류 레이어



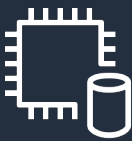
세부 분류 및 예시



Source : <https://a16z.com/2023/01/19/who-owns-the-generative-ai-platform>

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

생성 AI 코호트 Cohort



Model Provider (Builder)

자체 파운데이션
모델 구축



Model Tuner

타 모델빌더의
파운데이션 모델에
자사 데이터로 도메인
전문화/튜닝된 모델을
만들어 자사 목적에
맞게 활용



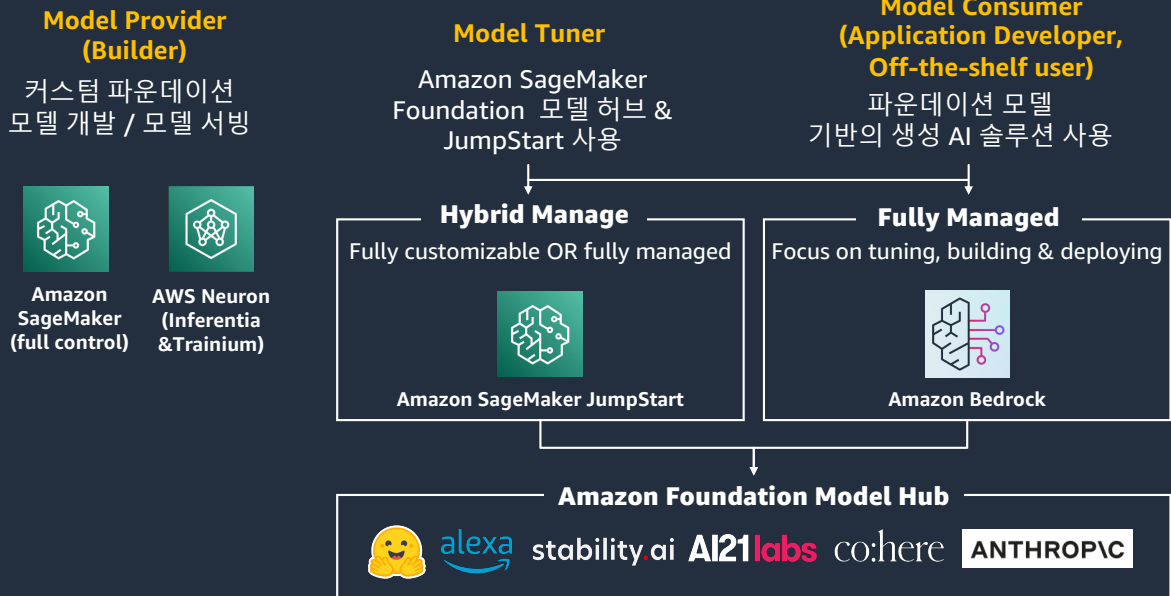
Model Consumer (Application Developer, Off-the-shelf user)

타 모델빌더의
파운데이션 모델을
튜닝 없이 API로
호출하여 자사 AI
애플리케이션 개발

aws

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

AWS의 생성 AI 전략



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

유니콘이 된 AI 스타트업

회사	차별화 요소	주요 영역	기업 가치	프로덕트
Hugging Face	Best for Community-Driven AI Development	<ul style="list-style-type: none"> Generative AI forums Open-source AI model development 	Private company; estimated at around \$2 billion.	<ul style="list-style-type: none"> BLOOM Expert Acceleration Program Private Hub Inference Endpoints AutoTrain Hardware
Cohere	Best for Natural Language Processing	<ul style="list-style-type: none"> Natural language processing and modeling Enterprise conversational AI 	Private company; estimated at around \$6 billion.	<ul style="list-style-type: none"> Embed Neural Search Summarize Generate Classify
Anthropic	Best for Customizable Content Generation	<ul style="list-style-type: none"> Content generation and editing Text classification and summarization Customer support 	Private company; estimated at around \$5 billion.	<ul style="list-style-type: none"> Claude Claude Instant
Jasper	Best for Marketers	<ul style="list-style-type: none"> Content generation Language modeling 	Private company; estimated at around \$1.5 billion.	<ul style="list-style-type: none"> Jasper Art Jasper Chat
Stability AI	Best Foundational Model for Other Generative AI Solutions	<ul style="list-style-type: none"> Content generation and editing Open-source AI app development 	Private company; estimated at around \$1 billion.	<ul style="list-style-type: none"> Stable Diffusion 2.0 Stable Diffusion Reimagine DreamStudio Photoshop Plugin Blender Plugin Platform API

Source: <https://www.eweek.com/artificial-intelligence/generative-ai-companies/>

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

AI 스타트업의 파운데이션 모델 on AWS

AI21labs

Jurassic-2

텍스트 생성을 위한 다국어 LLM
(Spanish, French, German,
Portuguese, Italian, and Dutch)

ANTHROPIC

Claude

대화 및 텍스트 처리 작업을 위한
LLM

stability.ai

Stable Diffusion

독창적이고 사실적인 고품질
이미지, 작품, 로고, 디자인 생성

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

한국 초거대 AI 모델 기업

- 생성 AI 생태계가 형성되고 있는 초기 단계
- 클라우드 인프라 확보 및 파트너십

회사	프로덕트	특징
네이버	하이퍼클로바	<ul style="list-style-type: none"> • 2,040억개의 매개변수 • 한국어 집중 훈련 (훈련 데이터 중 97%)
카카오브레인	KoGPT	<ul style="list-style-type: none"> • 한국어 특화 언어 모델 • 구글 TPU 활용, 연산 속도 고도화
LG AI연구원	엑사원	<ul style="list-style-type: none"> • 언어, 이미지 등을 다루는 멀티모달 모델 • 제조, 금융, 제조 등에 1% 전문가 목표
KT	믿음	<ul style="list-style-type: none"> • (B2B) 초거대 AI 제작 도구 믿음 렛츠 제공 • (B2C) 지니TV 음성대화 서비스
SKT	에이닷	<ul style="list-style-type: none"> • 사람과 대화하듯 자연스러운 소통이 장점 • 웨이브, 플로 콘텐츠 추천 등 친구처럼 일상의 디지털 메이트가 목표

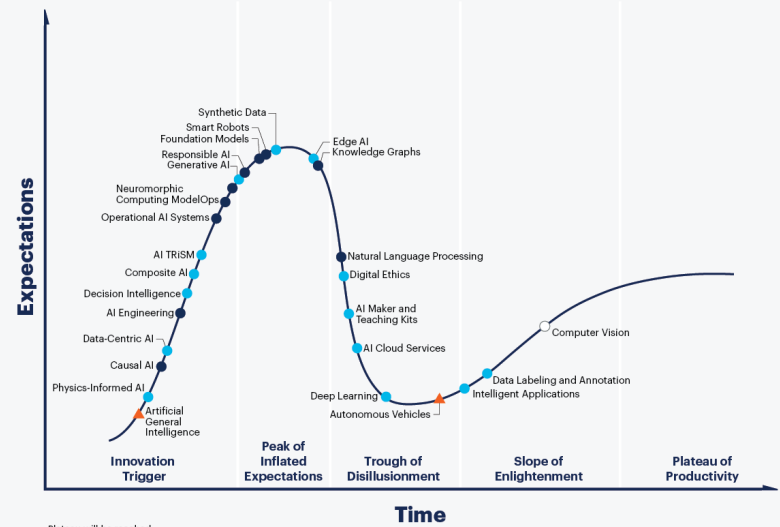
Source: 한국지능정보사회진흥원. 현대 인공지능의 역사적 사건 및 산업·사회 변화 분석

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

인공지능 성숙도 하이프 사이클 (출처: 가트너)

Source: <https://www.gartner.com/en/articles/whats-new-in-artificial-intelligence-from-the-2022-gartner-hype-cycle>

Hype Cycle for Artificial Intelligence, 2022



gartner.com

Source: Gartner
© 2022 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner and Hype Cycle are registered trademarks of Gartner, Inc. and its affiliates in the U.S. 1957302

As of July 2022

Gartner

English *all*-cleft constructions

A corpus-based approach

Youn-Gyu Park and Jong-Bok Kim (KYUNG HEE UNIVERSITY)

youngyu0406@khu.ac.kr, jongbok@khu.ac.kr

2023 KACL-KASELL Spring Conference
June 3, 2023



Introduction

- 1 / 24

- 2 / 24

Key properties

Syntactic property

English *all*-cleft constructions: A corpus-based approach

Filler-gap dependency

- Specificational *all*-cleft constructions **usually display the filler-gap dependency** between a syntactic gap in an *all*-cleft and a post-copula, focus-bearing 'pivot' (Kay 2010, a.o.):
 - (4) a. All that John ate ______i for lunch was *a banana*_i. (Homer 2019: 1)
b. John ate *a banana* for lunch.
 - However, the filler-gap dependency gets noticeably weaker in some specificational uses, or even disappears in predicational ones (Homer 2019; Tellings 2020).
 - (5) a. All they could do ______i was *embrace each other*_{*i}. (Tellings 2020: 4, adapted)
b. ?? They could do *embrace each other*.
 - (6) a. All that John ate ______i for lunch was *tasty*_{*i}. (Homer 2019: 4)
b. * John ate *tasty* for lunch.

3 / 24

Smallness reading

- In addition to the exhaustiveness, the *all*-cleft construction displays **scalar readings**, presupposing some sort of scalar model (Fauconnier 1975; Kay 2013; Tellings 2020).

(7) a. All I can eat is HALF *a pizza*. (Kay 2013:4) [quantity-scale]
 b. All he is is *a SIMPLE employee*. (Tellings 2020:8) [fame-scale]
- In addition, the scalar reading induces the smallness effect under some occasions, implying that the denoted entity “is not much” (Homer 2019; Tellings 2020).

(8) All John ate for lunch was a banana and a strawberry. (Homer 2019: 2)
 (⇒ A banana and a strawberry *aren't much* for lunch)

4 / 24

Previous analyses

A derivational approach

- Tellings (2020) argues that the *all*-cleft is a phonological realization of a relativization of *only*. This can account for the exhaustiveness of the constructions (Tellings 2020: 21):
 - (9) a. Ed **only** eats a salad. [non-clefted counterpart]
 - b. [_{FoCP} [a salad]_i Foc⁰ [Ed only eats *t_i*]] [movement to Spec, FocP]
 - c. [_{FORCEP} all_m [_{TOPP} [(that/what_{*t_i*}) Ed ~~only~~_m eats *t_i*]_j Top⁰ [_{FoCP} [a salad]_i Foc⁰ *t_j*]]] **[all-relativization]**
 - d. [_{FORCEP} all_m [_{TOPP} [(that/what_{*t_i*}) Ed ~~only~~_m eats *t_i*]_j is [_{FoCP} [a salad]_i Foc⁰ *t_j*]]] [copula insertion]
- However, since the analysis presumes the filler-gap dependency between pre- and post-copula entities, it can cover limited cases of the specificational *all*-clefts:
 - (10) a. [All that one has to do _____] is to start training earlier. (Kay 2010: 4)
 - b. * One only has to do to start training earlier.
 - (11) a. [All that John ate _____ for lunch] was *tasty*. (Homer 2019: 4)
 - b. * John only ate *tasty* for lunch.

5 / 24

A deletion-based approach of pseudocleft constructions

- A deletion-base approach on pseudocleft constructions is sketched by Ross (1972). The analysis is based on the deletion of objects of overlapping materials:
 - (12) a. What John read was *novels*. (den Dikken 2006:381)
 - b. John read novels.
 - c. [[*What* John read ____] [was [~~John read novels~~]]]
- This one can only cover restrictive instances of the specificational *all*-clefts, as the derivational analysis can (cf. Tellings 2020).
 - (13) All that one has to do is to start training earlier. (Kay 2010: 4)
 - a. * One has to do to start training earlier.
 - b. [[*All* that one has to do ____] [is [~~*all that one has to do~~ to start training earlier]]].
 - (14) All that John ate for lunch was *tasty*. (Homer 2019: 4)
 - a. * John ate *tasty* for lunch.
 - b. [[*All* that John ate ____ for lunch] [was [~~*John ate tasty~~]]].

6 / 24

Corpus investigation

Methodology

- To propose a plausible, robust analysis of both usages of *all*-clefts, we conducted a corpus investigation with authentic usages from the COCA.¹

(15) **Search strings:**

- a. PUNC all PRON (500)
- b. PUNC all that (500)

- Then, irrelevant tokens are manually filtered-out from the initial dataset:

- (16) a. **All compares are** done using the effective address, [...] . (COCA 1994 ACAD)
b. **All that** other stuff is secondary. (COCA 1999 NEWS)
c. This thrum, it is mine, and we beat together : **All worry is** a desperate wanting of my own way. (COCA 2012 BLOG)

¹COCA stands for the Corpus of Contemporary American English (Davis 2008-), one of the members of the BYU corpora family. It contains 1000+ million instances of English sentences.

Dataset and variables

- The final dataset contains 597 instances of *all*-cleft constructions.

Type	Initial dataset	Filtered token	Final dataset
PUNC all PRON	500	53	447
PUNC all that	500	350	150
Total	1,000	403	597

Table 1: Search strings and final dataset

- The following variables are adopted in analyzing the dataset to see the linguistic patterns of the construction:
 - (17) **Variables:**
 - a. Syntactic categories of post-copula elements: NP, AP, VP_[bse], VP_[ing], CP, S, ...
 - b. Filler-gap dependency: Violation / No violation
 - c. Reading of *all*-clefts: Specificational / Predicational
- Additionally, we conduct a co-varying collexeme analysis to see the overall constructional properties.

8 / 24

Syntactic categories of post-copula elements

- Various types of post-copula elements were found during the investigation:
 - (18) All I need **is** an opening, a moment of weakness. (NP, COCA 2008 FIC)
 - (19) All that we see **is** *noble and good and pure*. (AP, COCA 2010 MOV)
 - (20) a. All that did **was** make everyone angrier. (VP_[bse], COCA 1997 SPOK)
 - b. All that's the latest **is** covered 24 hours a day on CNN. (VP_[en], COCA 1991 SPOK)
 - c. All I wanted **was** to be with Bethany. (VP_[to-inf], COCA 1996 MAG)
 - (21) All I want to know **is** how he found out. (*wh*-cl, COCA 2003 TV)
 - (22) All I can remember **is** my parents had nothing. (S, COCA 2010 NEWS)
 - (23) All that remains **is** for you to open the gates. (CP, COCA 2019 MOV)
- The most frequent post-copula element was NP, followed by VP_[bse], CP_[that] and so forth.

Filler	NP	AP	VP				<i>wh</i> -CLAUSE	S	CP		Total
			bse	en	ing	to-inf			for	that	
Freq.	214	2	170	5	22	28	15	67	9	65	597

Table 2: Distribution of syntactic categories of post-copula elements (Raw freq.)

9 / 24

The filler-gap dependency

- The dataset indicates that, frequently, the filler-gap dependency is preserved (cf. Tellings 2020).

- (24) a. All you need to add is [_{NP} *a little milk*]. (COCA 2013 MAG)
b. You only need to add a little milk.

- Still, there are instances of *all*-cleft constructions with the filler-gap discrepancy.

- (25) a. All I've done is [_{VP[base]} *take off your sock*]. (COCA 2015 MOV)
b. * I only have done *take off your sock*.

- (26) a. All he's doing is [_{VP[ing]} *being a cheerleader for them*]. (COCA 1997 NEWS)
b. * He is only doing *being a cheerleader for him*.

Syn.Cat	AP	NP	wh-cl	S	CP	VP _[base]	VP _[en]	VP _[ing]	VP _[to-inf]	Total
No violation	-	214	15	67	74	-	-	15	22	407
Violation	2	-	-	-	-	170	5	7	6	190

Table 3: The filler-gap dependency between cleft-clauses and post-copula elements (Raw freq.)

10 / 24

Specificational and predicational *all*-cleft constructions

- The dataset indicates the construction is usually construed in its specificational reading.

- (27) a. All that remained of any use was *the desk*. (COCA 2017 FIC)
b. All he's doing is *being a cheerleader for them*. (COCA 1997 NEWS)

- The data indicates the construction can be used in its predicational usage, even if not so frequently.

- (28) a. All that we see is *noble and good and pure*. (COCA 2010 MOV)
b. All that's the latest is *covered 24 hours a day on CNN*. (COCA 1991 SPOK)

Type	Specificational	Predicational	Total
Raw Freq.	590	7	597

Table 4: Distribution of the predicational and specificational copula (raw freq.)

11 / 24

A co-varying collexeme analysis

- We conducted a co-varying collexeme analysis to investigate whether there are **strong attraction/repulsion relations** between gap-licensing verbs and post-copula elements.
- (29) **Collexeme analysis:**²
- “to quantify how much words in one slot of a construction are attracted to or repelled by words in a second slot of the same construction” (Gries 2019)
- (30) a. All I know is that the ink is 600 years old. (COCA 2003 MOV)
- b. [All I know] is [that the ink is 600 years old].
- ATTRACTION / REPUSSION
- The dataset yields 79 types of verbs. To get a clear result, the lexical items with low frequency (less than 3) in the dataset were excluded from the analysis (See Table 5).

²The analysis was conducted on the R (ver. 4.2.2) environment on the Rstudio (ver. 2023.3.0.386), using the package ‘Coll.analysis 4.0’.

Gap licensing verbs

Rank.	V_lex	Freq.	Rank.	V_lex	Freq.	Rank.	V_lex	Freq.	Rank.	V_lex	Freq.
1	do	188	2	know	51	3	say	43	4	need	39
5	want	23	6	matter	22	-	remain	22	8	leave	18
9	see	13	9	take	13	11	hear	11	11	have	11
13	care.about	10	14	be	8	14	get	8	16	require	7
17	tell	6	17	be.missing	6	17	happen	6	20	remember	5
21	ask	4	21	mean	4	21	think	4	21	find	4
21	think.about	4	26	think.of	3	26	talk.about	3	26	ask.for	3
29	hope.for	2	29	change	2	29	accomplish	2	29	bring	2
29	be.important	2	29	come.out	2	29	be.necessary	2	36	learn	1
36	talk	1	36	agree.on	1	36	assume	1	36	expect	1
36	deserve	1	36	act.on	1	36	saw	1	36	become	1
36	seem	1	36	seek	1	36	seen	1	36	address	1
36	read	1	36	write	1	36	claim	1	36	present	1
36	carry.out	1	36	give	1	36	offer	1	36	be.sure.of	1
36	go.by	1	36	concern	1	36	hold.on	1	36	love	1
36	imagine	1	36	suggest	1	36	be.good.for	1	36	distinguish	1
36	legitimate	1	36	coach	1	36	achieve	1	36	greet	1
36	go	1	36	come	1	36	glitter	1	36	be.holy	1
36	prove	1	36	separate	1	36	keep	1	36	stand	1
36	teach	1	36	endure	1	36	list	1			

Table 5: Distribution of gap licensing verbs and their frequency (Raw freq.)

A co-varying collexeme analysis: Result

- The result indicates that the verb lexeme *do* displays the most strong attraction with the post-copula element whose form is the VP_[bse].³

(31) a. All you gotta *do* is *say yes*. (VP_[bse], COCA 2012 TV)
b. All he's *doing* is being a cheerleader for them. (VP_[bse], COCA 1997 NEWS)
- Meanwhile, the verb lexeme *do* repels other types of post-copula elements, such as:

(32) a. All I had *done* is clicked on the group page and then clicked away. (VP_[en], COCA 2012 WEB)
b. If I stuck the plug in the socket, all I had to *do* was to throw the lever. (VP_[to-inf], COCA 1997 FIC)
- The Coll.str. is followed by other combinations such as *say-S* (28 tokens, LLR = 76.90), *need-NP* (36 tokens, 65.55), and so forth.

(33) a. All I can *say* is *thank you for appreciating all my hard work*. (COCA 2008 FIC)
b. All you *needed* was *a good memory*, that's all. (COCA 1993 FIC)

³Verb lexeme *do*; Collocational strength (Coll.str.) = Logged Likelihood Ratio (LLR) = 553.30. See Table 6.

A co-varying collexeme analysis: Result (Cont'd)

<i>W</i> ₁ (pre.cop.)	<i>W</i> ₂ (post.cop.)	<i>W</i> ₁ x <i>W</i> ₂ .freq.	<i>W</i> ₁ .freq.	<i>W</i> ₂ .freq.	relation	LLR (Coll.str.)
<i>do</i>	V.bse	170	188	170	attraction	553.29667
	V.ing	7	188	18	attraction	0.12973
	V.en	1	188	3	repulsion	-0.00319
	V.to-inf	6	188	23	repulsion	-0.85534
<i>know</i>	NP	4	188	179	repulsion	-159.93242
	CP.that	24	51	57	attraction	51.83528
	wh-cl	5	51	12	attraction	9.00864
	S	21	51	66	attraction	31.42489
<i>say</i>	NP	1	51	179	attraction	-35.03238
	S	28	43	66	attraction	76.90187
	V.en	1	43	3	attraction	1.58528
	CP.that	10	43	57	attraction	6.32251
<i>need</i>	NP	4	43	179	repulsion	-14.63629
	NP	36	39	179	attraction	65.55040
	CP.that	2	39	9	attraction	2.07311
	VP.to-inf	1	39	23	repulsion	-0.34283
<i>want</i>	VP.to-inf	8	23	23	attraction	24.67694
...						

Table 6: The results of the co-varying collexeme analysis on the dataset

Data discussion

Two types of *all*-cleft constructions

English *all*-cleft constructions: A corpus-based approach

***All* in its specificational, exhaustiveness usage**

- The corpus research yields both specificational and predicational *all*-cleft constructions. The observation parallels the properties of pseudo-clefts (Higgins 1979).
- Granting the observation, we can sketch the semantic representations and their information-structure as the followings (cf. Higgins 1979; Lambrecht 2001):

(34) All I need is an address. (= 'I need *nothing but an address*')

a. $\iota x[\text{need}(i, x)]$: *an address*

b. Info-st

Presupposition:	'I need something <i>x</i> '
Focus:	' <i>an address</i> '
Assertion:	' <i>x = an address</i> '

- The exhaustiveness reading that is induced by the construction can be notated with the **iota operator**, as shown in (34a).

***All* in its predicational, universal quantification usage**

- However, things are different in the predicational *all*-cleft constructions, since it serves as a “pseudo-adjective” (Lambrecht 2001).

(35) All that we see is noble and good and pure.
 (= ‘We see something and all of them are noble and good and pure.’)

 - a. $\forall x[\text{see}(\text{we}, x) \ \& \ \text{noble}(x) \ \& \ \text{good}(x) \ \& \ \text{pure}(x)]$
 - b. Info-st

Presupposition: ‘We see a set of *x*’
 Predicate: ‘All *x* is noble, good, and pure’
- Unlike *all* in the specificational one, it seems that *all* here functions as a **universal quantifier**, and the post-copula element serves as a predicate, as shown in (35a).
- This assumption concords with the function of other predicational pseudoclefts (cf. Higgins 1979; Lambrecht 2001; Kim To appear).

17 / 24

Syntactic patterns of verb lexeme *do* in *all*-cleft

- During the investigation, we figured out that the verb lexeme *do* in *all*-clefts with a VP pivot only appears in specificational *all*-clefts.

(36) “All we can do is relax and get it going,” said manager Art Howe. (COCA 1992 NEWS)
 $\Rightarrow \iota x[\Diamond \text{do}(\text{we}, x)]: \text{relax and get it going}$
- We assume that this is the realization of the constructional interaction between the specificational *all*-clefts and **the *do-be* constructions** (Flickinger and Wasow 2013: 2):

(37) a. _[THING-CL] The thing I’m **doing** is trying to learn from my mistakes.
 b. _[ALL-CL] All the government **does** is send out checks.
 c. _[WH-CL] What you have to **do** is get ready.

18 / 24

Syntactic patterns of verb lexeme *do* in *all*-cleft

- We observed three syntactic patterns, also noticed by Flickinger and Wasow (2013).

Pattern 1:

If The embedded *do* is in its finite form, the VP pivot should be in its *bse* form.

- (38) All that [_{VP[fin]} did] was [_{VP[bse]} make everyone angrier]. (COCA 1997 SPOK)

Pattern 2 & 3:

If the embedded *do* is in its nonfinite form, the VP pivot should be...

- either in its *bse* form;

- (39) Meantime, all she can [_{VP[bse]} do] is [_{VP[bse]} maintain hope]. (COCA 1992 NEWS)

- or its form follows the form of pre-copula *do*;

- (40) All we're [_{VP[ing]} doing] now is [_{VP[ing]} locking everybody up]. (COCA 1993 ACAD)

19 / 24

Contrastive focus inducing smallness reading

- One of the unique semantic properties of the specificational *all*-cleft construction is the smallness effect, implying that something “*isn't much*” (Kay 2013; Tellings 2020).
- The dataset implies that a pivot requires a counterpart material that receives **contrastive focus** to induce the smallness reading:

(41) a. All I got is [_{FOC} *the garage*].
 (= ‘I got nothing more than the garage’ \nRightarrow The garage isn't much)

b. All I got is [_{FOC} *the garage*]. You got [_{CONT.FOC} **THE WHOLE REST OF THE DAMN HOUSE**].
 (= ‘I got nothing more than the garage’ \Rightarrow The garage isn't much)
- Granting this observation, we assume that the contrastive focus is one of the factors that have to do with inducing the smallness readings of the construction.

20 / 24

Derivation-based and deletion-based analyses

- The corpus investigation shows that the most salient construct of the construction is...
(42) All we can do is try to do our best. (COCA 1990 FIC)
- Both the derivation-based and the deletion-based views cannot analyze such cases (cf. Ross 1972; Tellings 2020).
(43) a. * We can **only** do try to do our best.
[All we can do _____i] is [try to do our best]_i
b. * We can do try to do our best.
[All we can do ____] [is [~~*we can do~~ try to do our best]]
- Furthermore, they cannot account for the predicational ones either.
(44) All we see is noble. (COCA 2010 MOV, adapted)
a. * We **only** see noble.
[All we see _____i] is [noble]_i
b. * We see noble.
[All we see ____] [is [~~*We see~~ noble]]

21 / 24

Two types of copula and two types of *all*-clefts

- As an alternative, we propose a construction-based approach to *all*-cleft constructions. We assume that it is the copula constructions that decide the types of *all*-clefts.
(45) a. Specificational Copula (\uparrow *cop-lxm*)

$$\begin{bmatrix} \text{spec-cop} \\ \text{ARG-ST} & \langle \text{XP}[\iota x[\text{P}(x)]], \text{YP}_y \rangle \\ \text{SEM} & \text{identify_rel}(x, y) \end{bmatrix}$$
b. Predicational Copula (\uparrow *cop-lxm*)

$$\begin{bmatrix} \text{pred-cop} \\ \text{ARG-ST} & \langle \text{XP}, \text{YP} \rangle \\ \text{SEM} & [\text{YP}(\text{XP})] \end{bmatrix}$$
- Given the two types of copula constructions, we can analyze the copulas in (46) as the followings, respectively:
(46) a. All you need **is** love.

$$\begin{bmatrix} \text{spec-cop} \\ \text{ARG-ST} & \langle \text{NP}[\iota x[\text{need}(\text{you}, x)], \text{NP}_y] \\ \text{SEM} & \text{identify_rel}(x, y) \end{bmatrix}$$
b. All you need **is** in this box.

$$\begin{bmatrix} \text{pred-cop} \\ \text{ARG-ST} & \langle \text{NP}_{[\text{all}]}, \text{PP}_{[\text{in}]} \rangle \\ \text{SEM} & \forall x[\text{need}(\text{you}, x) \ \& \ \text{in}(x, \text{box})] \end{bmatrix}$$

22 / 24

Advantages

- The non-derivational direct-licensing approach can take the following advantages:

Construction-based approach:

The analysis can account for the constructional properties of the *all*-cleft constructions, that are shared with the other members of the pseudocleft construction family.

- (47) a. What you have to do is get ready. (Flickinger and Wasow 2013: 2)
b. All you have to do is get ready.

Non-derivational approach:

The analysis can cover not only the specificational *all*-clefts with canonical pivots, but also idiosyncratic VP pivots.

- (48) a. All we can do is try to do our best. (COCA 1990 FIC)
b. All we can do is not even close to our goals.

23 / 24

Summary

Summary

- We started by introducing the linguistic properties of English *all*-cleft constructions and some theoretical issues that the previous literature holds.
- To see linguistic patterns of the construction, we performed a corpus investigation accompanying a co-varying collexeme analysis. The dataset indicates that there are cases the previous literature cannot properly account for.
- Based on the observation, we briefly sketched theoretical implications and basic analysis of the *all*-cleft constructions. The analysis can cover not only linguistic properties of the construction but also constructional properties shared with the other members of the pseudocleft construction family.

24 / 24

References

- Bock, Kathryn, and Carol A. Miller. 1991. Broken agreement. *Cognitive Psychology* 23(1): 45-93.
- Boeckx, Cedric. 2007. Pseudoclefts: A fully derivational account. In M.T. Hany Babu Josef Bayer and Tanmoy Bhattacharya (eds.), *Linguistic theory and south Asian languages: Essays in honour of K. A. Jayaseelan*. 29-40. Amsterdam: John Benjamins Publishing Company.
- Cann, Ronnie. 1993. *Formal semantics: An introduction*. Cambridge: Cambridge University Press.
- Chomsky, Noam. 1972. Deep structure, surface structure and semantic interpretation. In Aus dem Buch (ed.), *Studies on Semantics in Generative Grammar*. 62-119. Boston: De Gruyter Mouton.
- Collins, Peter. 1991. *Cleft and pseudo-cleft constructions in English*. New York, NY: Routledge.
- Davies, Mark. 2008-. *The Corpus of Contemporary American English (COCA)*. Available online at <https://www.english-corpora.org/coca/>.
- Delahunty, Gerald P. 1984. The analysis of English cleft sentences. *Linguistic Analysis* 13(2): 63-113.

References

- den Dikken, Marcel. 2006. Specificational copular sentences and pseudoclefts. In Martin Everaert and Henk van Riemsdijk (eds.), *The Blackwell Companion to Syntax*. Malden, MA: Blackwell.
- Erteschik-Shir, Nomi. 1986. *Wh*-questions and focus. *Linguistics and Philosophy* 9(2): 117-149.
- Fauconnier, Gilles. 1975. Pragmatic scales and logical structure. *Linguistic Inquiry* 5(3): 353-375.
- Flickinger, Dan and Thomas Wasow. 2013. A corpus-driven analysis of the *do-be* construction. In Philip Hofmeister and Elisabeth Norcliffe (eds.), *The core and the periphery: Standing on the shoulder of Ivan A. Sag*. 35-63. Stanford, CA: CSLI publications.
- Goldberg, Adele. 2006. *Constructions at work: The nature of generalization in language*. New York, NY: Oxford University Press.
- Gries, Stefan Th. 2019. 15 years of collocations: Some long overdue additions/corrections (to/of actually all sorts of corpus-linguistics measures). *International Journal of Corpus Linguistics* 24(3): 385-412.
- Gries, Stefan Th. 2022. *Coll.analysis 4.0. A script for R to compute perform collocational analyses*. Available online at <https://www.stgries.info/teaching/groningen/index.html>.

References

- Higgins, Francis Roger. 2015. *The pseudo-cleft construction in English*. New York, NY: Routledge.
- Homer, Vicent. 2019. That's all. In Richard Stockwell, Maura O'Leary, Zhongshi Xu and Zhao .L. Zhou (eds.), *Proceedings of the 36th West Coast Conference on Formal Linguistics*, 1-21. Somerville, MA: Cascadia Proceedings Project.
- Kay, Paul. 2013. The limits of (construction) grammar. In Thomas Hoffmann and Graeme Trousdale (eds.), *The Oxford handbook of construction grammar*, 32-48. Oxford: Oxford University Press.
- Kim, Jong-Bok. To appear. *Cleft constructions*. Seoul: Kyung Hee University. Ms.
- Kim, Jong-Bok and Laura Michaelis. 2020. *Syntactic constructions in English*. Cambridge: Cambridge University Press.
- Kim, Jong-Bok and Peter Sells. 2007. *English Syntax: An Introduction*. Stanford, CA: CSLI publications.
- Lambrecht, Knud. 2001. A framework for the analysis of cleft constructions. *Linguistics* 39(3): 463-516.
- Pollard, Carl, and Ivan A. Sag. 1994. *Head-driven phrase structure grammar*. Chicago, IL: University of Chicago Press.

References

- Posit team (2023). *RStudio: Integrated Development Environment for R*. Posit Software, PBC, Boston, MA. <http://www.posit.co/>.
- Preacher, Kristopher J. and Nancy E. Briggs (2001, May). *Calculation for Fisher's Exact Test: An interactive calculation tool for Fisher's exact probability test for 2 x 2 tables [Computer software]*. <http://quantpsy.org>.
- R Core Team (2017). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Ross, John Robert. 1972. Act. In Donald Davidson and Gilbert Harman (eds.), *Semantics of Natural Language*. 70-126. Dordrecht: Reidel.
- Sag, Ivan. A. 2008. *English filler-gap construction*. Manuscript: Stanford, CA: Stanford University.
- Stefanowitsch, Anatol and Gries, Stefan. Th. 2003. Collostructions: Investigating the interaction of words and constructions. *International journal of corpus linguistics* 8(2): 209-243.
- Tellings, Jos. 2020. An analysis of *all*-clefts. *Glossa: a journal of general linguistics* 5(1): 1-25.

합성어 ‘국뽕’의 의미 해석과 은유

나익주 (전남대학교)

손흥민 선수가 영국프로축구 리그에서 대단한 활약을 하던 몇 년 전부터 유럽에서 활동하는 우리나라 축구 선수들을 주제로 다루는 유튜브를 자주 보곤 했다. 며칠 전에는 김민재 선수의 대활약을 절절히 보도하는 유튜브를 보고 있다가 친구로부터 편지를 들었다. “또 국뽕이냐?”

그 순간 처음 들었던 어구의 의미를 이해하지 못해 잠깐 당황했던 오래 전의 한 장면이 떠올랐다. 바로 ‘뽕(을) 맛다’라는 어구였다. 회의 내내 하품만 하다가 회의실을 나오는데 가까운 후배가 “선생님, 뽕 맛은 거 같아요.”라고 말을 건넸다. “어? 무슨 말이야?”라고 되물으니, “히로뽕 맛았냐구요. 오늘 좀 이상해서요.”라고 다시 말했다. 그러자 그의 의도를 바로 알아차리며 “어 어제 좀 술을 좀 많이 마셨어.”라고 대답했다.

‘뽕’의 의미적 다중성: 동음이의적? 다의적?

『연세한국어사전』에는 ‘뽕’은 두 개의 표제어로 실려 있다. 하나(뽕1)는 ‘뽕잎’과 ‘뽕나무’의 준말이라고 설명이 달려 있는 명사이며, 다른 하나(뽕2)는 “막혀 있던 기체나 가스가 좁은 구멍으로 갑자기 터져 나오는 소리”나 “작은 구멍이 또렷하게 뚫리는 소리나 모양”을 가리키는 부사이다. 『표준국어대사전』과 『고려대학교 한국어대사전』에는 “필로폰’을 속되게 이르는 말”이라는 설명을 달고 있는 또 하나의 표제어(뽕3)가 실려 있다. 『고려대학교 한국어대사전』에는 민속학에서 사용하는 용어인 “그네뛰기에서 그네를 타고 올라 발끝으로 칠 수 있도록 높이 달아 놓은 방울”이라는 설명을 달고 있는 표제어가 하나 더 실려 있다. 국립국어원의 우리말샘에는 “치레로 물건 바닥 한복판에 박아 넣는 다른 물건”을 가리키는 낱말 ‘뽕’의 속된 말인 ‘뽕’도 표제어로서 실려 있다.

속된 말 표제어와 민속학의 표제어를 제외하면, 일상적으로 사용되는 ‘뽕’의 표제어는 3개이다. 이 세 가지 표제어 ‘뽕’은 각각 다음과 같은 문장들에서 사용된다.

- 아낙네들이 뽕을 따 머리에 이고 내려오고 있었다. (뽕1)
- 그는 방귀를 뽕 뼀 뒤 무색한지 얼굴을 붉혔다. (뽕2)
- 뽕을 계속 맞으면서 그는 육체적으로 점점 망가지기 시작했다. (뽕3)

동일한 소리의 낱말이 다수의 의미를 지니고 있을 때, 이 낱말은 동음이의 현상의 사례일 수도 있고 다의 현상의 사례일 수도 있다. 동음이의어와 다의어를 가르는 가장 중요한 기준은 어떤 낱말이 지닌 다수의 의미들 사이에 어떤 관련성이 지각되는가의 여부이다. 위의 예문에서 보듯이 ‘뽕’의 세 의미 사이에서는 어떤 의미적 관련성도 지각할 수 없다. 따라서 ‘뽕’은 한 낱말이 아니라, 우연히 동일한 소리(/bbɔŋ/)를 지니고 있는 별개의 세 낱말(뽕1, 뽕2, 뽕3)이다. 당연히 ‘뽕’은 동음이의 현상의 사례이다. 한국어 사전에서도 이러한 인식을 반영하여 ‘뽕’을 하나의 표제어 아래에 다수의 의미를 제시하는 것이 아니라 다수의 표제어로 나누어 각 의미를 제시하였다.

‘국뽕’의 의미: “국가가 소유한 ‘히로뽕’인가”?

낱말 ‘국뽕’은 당연히 ‘국가’의 ‘국(國)’과 ‘뽕’이 결합한 합성어이다. 그러면 ‘국뽕’의 ‘뽕’은 이 표제어들 중 어느 ‘뽕’일까? “방구”의 ‘뽕’? ‘뽕나무’의 ‘뽕’? “필로폰”의 ‘뽕’? 아직은 ‘국뽕’이 한국어사전에 정식으로 표제어로서 실려 있지 않으며, 네이버 오픈 사전에만 ‘국가’의 ‘국’과 ‘히로뽕’의 ‘뽕’이 합쳐진 말로서 “국수주의 민족주의가 심하며 타민족에 배타적이고 자국만이 최고라고 여기는 행위나 사람”을 가리킨다는 설명이 있다. 그렇다면 ‘국뽕’의 ‘뽕’은 ‘히로뽕’이 ‘필로폰’의 일본식 발음이니까 당연히 “필로폰”의 ‘뽕’이다.

언론의 보도에서든 일상의 대화에서든 ‘국뽕’을 “국가가 소유한 히로뽕” “국가가 제작한

히로뽕” “국가가 관리하는 히로뽕” 등의 의미로 사용한 사례는 아직 찾아볼 수 없다. ‘국뽕’은 현재 독자적인 의미로는 사용되지 않으며, 다음 예문에서 보듯이 “지나치게 맹목적인 애국주의”라는 비유적 의미로만 사용된다.

- 언론은 국뽕에 대해 비판적인 태도를 취하지만 장사가 된다는 이유로 국뽕을 부추기는 데도 앞장을 선다. 김태익은 “신문 · 방송에는 ‘한국 ○○에 폭 빠진 외국인’ 기사가 줄을 잇는다”며 (한옥/자연/다산/퇴계/장아찌)
- “외국 박물관서 ‘일본해’ 지운 한국인”에 ‘국뽕’ 논란: 베트남의 한 박물관에 있다는 이 지구본에는…… ‘Sea of Japan(일본해)’이라 적혀 있던 것을 ‘Sea of Korea(한국해)’를 의미하는 것으로 보이는 ‘Sea of Coree’로 바꿔놓은 것이다. (* Coree는 Korea의 오기로 보임)
- 김완태 (아나운서) 편파중계 논란에 대해 누리꾼들은 “김완태 편파중계 논란, 국뽕이 심하셔요.” “김완태 편파중계 논란, 창피하네요.” “김완태 편파중계 논란, 해설의 질이 떨어지네요.” 등 다양한 반응을 보였다.

이러한 맥락이나 다른 많은 맥락에서 ‘국뽕’을 접했을 때 한국인들 중에서 이 낱말의 의미를 이해하지 못하는 사람은 거의 없다. 오히려 특별히 의식적으로 노력하지 않아도 듣는 순간 바로 그 의미를 알아차린다. 처음 들었을 때 ‘뽕 맛았어요?’의 은유적 의미를 바로 알아차리지 못했던 것은 사실 ‘뽕’이 ‘히로뽕’의 약어인줄 몰랐기 때문이지 나의 의미 처리 능력이 부족했기 때문이 아니었다. 원래 이 ‘뽕’은 비밀리에 마약을 제조하거나 밀수하거나 투여하는 사람들만이 사용하는 은어였다가 그 사용역이 일반인 화자들에게로 확대되었다.

‘국뽕’의 생성과 의미 처리: [편향적 생각은 유해 약물] 은유

‘국뽕에 취하다’ ‘국뽕을 제대로 맛다’ ‘국뽕을 부추기다’ 등의 어구를 듣는 순간, 우리 한국인들은 별다른 노력을 하지 않아도 이러한 어구의 의미를 “배타적인 민족주의와 심한 국수주의로 자국만이 최고라고 생각하며 다른 국가나 다른 민족을 폄하하는 언동을 거리낌 없이 하다”라고 바로 이해한다. 어떻게 이러한 즉각적인 의미 처리가 가능할까? 이것은 우리 한국인들의 마음속에 깊숙이 자리 잡은 개념적 은유인 [편향적 생각은 유해 약물] 덕택이다. 특정한 이론이나 사조만이 옳다고 믿는 [편향적 생각]을 진정제나 각성제, 흥분제, 특히 마약과 같은 [유해 약물]로 간주하는 이 은유적 사고는 인지적 무의식으로서 거의 자동적으로 작동한다. 그렇지 않다면, 우리는 ‘국뽕’이라는 낱말을 그렇게 쉽게 사용하지도 듣는 순간 곧장 그 의미를 이해하지도 못했을 것이다.

‘국뽕’이라는 낱말 이외에도 다른 국가의 이름 중 한 글자에 ‘뽕’을 더한 합성어도 현재 사용되고 있다. 아래의 기사 발췌문에서 보듯이 ‘일뽕’과 ‘중뽕’이 그러한 낱말이다. 물론 ‘일뽕’은 ‘일본’의 ‘일’과 ‘히로뽕’의 ‘뽕’이 결합한 합성어이고, ‘중뽕’은 ‘중국’의 ‘중’과 ‘히로뽕’의 ‘뽕’이 결합한 합성어이다. ‘일뽕’의 의미는 “일본이 정치, 경제, 문화 등 모든 측면에서 우월하다는 생각”을 가리키고, ‘중뽕’은 “중국이 그러하다고 생각하는 믿음”을 가리킬 것이다. 이 개념적 은유는 ‘국뽕’뿐만 ‘일뽕’이나 ‘중뽕’과 같은 어구의 생성과 의미 처리에도 작용하고 있다.

- 하루가 멀다하고 서로가 서로를 ‘국뽕’이니 ‘일뽕’이니 하며 공격하는 통에 요즘 온라인 세상은 평화로울 날이 없다.
- 중국이 서구적 질서에 긍정적인 대안을 제공해줄 것이라고 믿는 ‘중뽕’은 중국이 이룩해 낸 놀라운 성취나 아프리카의 인프라 투자 같은

최근에는 ‘돈’과 ‘히로뽕’의 ‘뽕’이 결합한 합성어 ‘돈뽕’도 사용하기 시작하고 있다. 이 낱말의 의미는 아래 기사 발췌문에서 보듯이 “사회 생활을 영위하는 데 다른 어떤 것보다도 돈이 더 중요한 가치를 지닌다고 믿는 경향”을 가리킨다.

- “이러다 재벌 되는 거 아냐?”: ‘나 일 왜 하고 있지?’ ‘돈뽕’은 이성과 감각을 마비시킨다. 코인 거래에 몰두하느라 하루 3~4시간 남짓 잠을 잤지만 피곤함이 느껴지지 않았다. ‘돈뽕’의 효력이 다하자 피로가 몰려왔다.
- 이익을 앞세우다 ‘돈뽕’에 중독된 경제파의 세상이 자본주의다. 정치가 사회를 집어삼킨 것이 국가 사회주의였다. ‘정치 우선주의’에 빠지면 권력만 강해진다. 권력을 앞세우다 ‘국뽕’에 취한 정치파의 세상이 망한 사회주의였다.

‘국뽕’과 ‘일뽕’ ‘중뽕’ ‘돈뽕’을 넘어서 히로뽕의 ‘뽕’은 이제 두 글자 낱말과도 결합해 합성어로 사용되기 시작했다. 예컨대 다음의 인터뷰 발췌문에 등장하는 ‘외국뽕’과 ‘예술뽕’이 그러한 사례이다.

- 20대엔 내가 남들과 다른 특별한 존재라고 생각했다. ‘외국뽕’ ‘예술뽕’에 차 있어서 ‘내가 외국에서 태어났으면 훨씬 잘 났을 텐데, 나는 가난해서 외국에 갈 수도 없어’라며 투덜대곤 했다.

위의 인터뷰에 등장하는 ‘외국뽕’과 ‘예술뽕’의 ‘뽕’은 어떤 것에 대한 ‘배타적인 우월주의 경향의 믿음’이 다소 희석된 ‘특별히 좋아하는 경향’을 가리킨다. 이러한 점에서 ‘국뽕’이나 ‘돈뽕’의 ‘뽕’과는 의미가 다소 다르다. 그렇지만 비록 배타적이지는 않더라도 무언가를 특별하게 좋아하는 성향은 ‘편향적 사고’의 약화된 사례이기 때문에 ‘외국뽕’과 ‘예술뽕’의 의미 처리에도 역시 [편향적 생각은 유해 약물]이라는 개념적 은유가 깔려 있다고 볼 수 있다.

왜 ‘국뽕’이라 명명했을까?: [부분이 전체를 대신함] 환유

유해약물에는 히로뽕 이외에도 대마초, 아편, 모르핀, 코카인, 헤로인, 졸피뎀, 프로포폴 등 다양한 품목이 있다. 그러면 ‘편향적 사고의 소유자’를 지칭하는 합성어를 만들 때 대마초의 ‘초’를 따서 ‘국초’라 하거나 아편의 ‘편’을 따서 ‘국편’이라 하지 않고 왜 하필 히로뽕의 ‘뽕’을 취해서 ‘국뽕’이라 했을까? ‘국초(國初/國礎)’나 ‘국편(國編)’ ‘국졸(國卒)’ ‘국모(國母)’ 등은 이미 다른 의미로 사용되고 있는 낱말이기 때문일까? 아니다. 현재 시점에서 ‘히로뽕’이 한국인들에게 ‘유해약물’, 특히 ‘마약’ 범주의 대표적 구성원으로 다가오기 때문으로 보인다. 물론 한국인들에게 ‘대마초’나 ‘아편’도 쉽게 ‘마약’의 전형적인 구성원으로 쉽게 떠오른다. 하지만 이 둘보다는 환각성이 더 강하다고 알려진 ‘히로뽕’이 편향적 사고 소유자의 비상식적인 언동이나 비합리적인 판단을 묘사하는 데 더 적절하다고 판단했을지 모른다.

실제로 어떤 종류의 약물을 복용하는지에 관계없이 상습적인 약물 복용자를 흔히 ‘뽕쟁이’이라 칭한다. 이러한 언어사용에서 볼 때, 한국인들에게는 ‘히로뽕’이 [마약] 범주의 원형적 구성원이라는 추정이 완전히 터무니없지는 않다. 편향적 사고 소유자를 지칭하기 위한 ‘국뽕’ 명명에는 [부분(히로뽕)이 전체(마약)를 대신함]이라는 개념적 환유도 작용하고 있다.

[지식 습득은 음식물 섭취] 은유

‘국뽕’이라는 합성어를 생성하고 그 의미를 처리하는 과정에 작용하는 마음속의 개념적 은유인 [편향적 생각은 유해 약물]은 홀로 유리되어 있지 않으며 다양한 개념적 은유와 관련이 있다. 우선 [균형잡힌 생각은 이로운 음식] 은유, [불온한 생각은 역겨운 음식] 은유, [흥미로운 생각은 맛있는 음식] 은유 등과 마찬가지로 [편향적 생각은 유해 약물] 은유는 [생각은 음식] 은유와 관련이 있으며, [(지식을) 학습하는 것은 (음식을) 섭취하는 것] 은유와도 관련이 있다. 열리고 펼쳐 있는 이 은유들과 관련이 있다는 점에서, 합성어 ‘국뽕’은 은유적으로 사용되는 아래와 같은 많은 표현들과 의미적으로 연결된다.

- 그렇게 뜯들이지 말고 네 생각을 빨리 말해라.
- 떠먹이는 주입식 교육에서 자기주도 학습으로 나아가야 해.
- 난 설익은 생각을 함부로 말하지 않아.

- 수업 그만 해요 소화불량 걸리겠어요.
- 그 연구는 몇 년의 숙성 과정을 거친 결과야.
- 독서는 사고의 양식이야.
- 그 이론은 천천히 곹씹어야 이해할 수 있어.
- 이는 보수 언론의 주장을 그대로 반영한 것으로 '국민의 의견'이라는 핑계를 대며 정권의 입맛에 맞게 역사·사회 교육을 흔든다는 비난을 면키 어렵다.

위의 예문은 한국인들의 마음속에 깊숙이 자리 잡고서 인지적 무의식으로 작동하는 [생각은 음식] 은유와 [지식 습득은 음식물 섭취] 은유의 언어적 발현이다. '국뽕'도 위의 표현들과 마찬가지로 [지식 습득은 음식물 섭취]이라는 개념적 은유에서 발현된 언어 표현이다. 물론 이 주장에는 '히로뽕'이 [마약] 범주의 구성원이자 '사람이 먹고 마시는 모든 것'을 가리키는 [음식물] 범주의 구성원일 수 없다는 반론이 나올 수 있다. 맞다. 분명히 '히로뽕'은 넓게는 [약] 범주, 좁게는 [마약] 범주의 구성원이다. 하지만 '음식이 보약이다.'나 '음식으로 못 고치는 병은 약으로도 못 고친다.'는 속설에서 알 수 있듯이, [약]과 [음식물]은 둘 다 상위 범주인 [섭취물]의 하위 범주로서 둘 사이에는 명확한 경계선이 없으며 오히려 일부 주변적인 구성원들은 중첩한다.

따라서 '히로뽕'은 분명히 [(마)약] 범주의 주변적인 구성원이지만, 동시에 [음식물] 범주의 주변적인 구성원일 수 있다. 의학적으로는 기준 한도 내에서 의료 목적으로 복용하면 신체에 유익할 수 있겠지만, 세속적으로는 아무리 적은 양의 마약도 몸에 흡수되기만 하면 그 자체로 신체적 건강을 파괴한다고 믿는다. 한마디로 마약은 신체 건강에 가장 유해한 섭취물이다.

'국뽕'을 비롯한 수많은 표현들에 깔려 있는 [지식 습득은 음식물 섭취] 은유는 근원 영역인 [음식물 섭취] 개념과 목표 영역인 [지식 습득] 개념 사이의 수많은 하위 대응으로 구성된다. 그 중의 하나가 바로 [생각은 음식] 은유이다. [음식물 섭취] 영역에는 음식물, 먹는 사람, 섭취 방식, 소화 등의 요소가 있다. 음식물에도 영양소가 풍부하고 몸에 유익한 음식물이 있는가 하면, 영양소는 거의 없고 몸에 해로운 영향을 끼치는 음식물도 있다. 한편 [지식 습득] 영역에는 학습하고자 하는 이론이나 사조와 학습자, 학습 방식, 이해 등의 요소가 있고, 이론이나 사조의 학습은 대체로 개념의 의미 파악이다.

구체적으로 [지식 습득은 음식물 섭취] 은유에서 이론이나 사조, 지식의 핵심 개념은 음식물에, 이 개념을 학습하는 사람은 음식물을 섭취하는 사람에, 반복 학습은 음식물의 곹씹기에, 개념의 완전한 이해는 음식물의 완전한 소화, 불완전한 이해는 소화 불량에 대응한다. 음식물의 섭취와 소화가 신체적 성장과 건강 유지에 필수적이듯이, 개념의 이해는 사고력 신장과 건강한 정신 활동에 필수적이다. 음식물에 이로운 음식물과 유해한 음식물이 있듯이, 지식에도 정신 건강에 좋은 지식도 있지만 유해한 지식도 있다. 특정한 음식물만을 가려서 즐겨 먹는 편식이 신체적인 건강 유지에 해롭듯이, 특정한 사조나 이념만을 학습하거나 신봉하는 지적 편향은 건전한 판단이나 균형 잡힌 사고를 가로막아서 정신 건강을 망치기 쉽다. 예컨대 히로뽕 복용자가 환청과 환각으로 인해 몸을 제대로 움직이지 못하듯이, 편향적인 사고의 소유자는 비상식적인 발언이나 행위를 하기 쉽다.

외연을 넓혀가는 'X뽕': 꼭 써야 할까?

개인의 애국심 표출이나 특정한 성향 그 자체를 비판할 생각은 없다. 그렇지만 다른 나라의 박물관에서 전시물에 적혀 있는 Sea of Japan을 Sea of Corea로 제멋대로 바꾸어 넣은 관람객의 행위나 올림픽 경기 생방송 중계에서 다른 나라 선수들의 실수를 폄하하는 아나운서의 발언은 일반인들의 상식으로 보아도 부적절하기에 비판받아 마땅하다. 그렇다고 해도 그들을 '마약복용자'에 비유하는 합성어 '국뽕'을 꼭 사용해야 했을까?

히로뽕의 '뽕'은 이제 '국뽕'과 '일뽕' '중뽕' '돈뽕'을 넘어서 '외국뽕' '예술뽕' '유세뽕' 등으로 그 외연을 넓혀 나가고 있다. 앞으로 어떤 '뽕' 합성어가 나올지 자못 궁금하다. 이러다가 특별한 기호를 지닌 사람들은 다 '아이돌뽕' '클래식뽕' '뽕짜뽕' '영화뽕' '극우뽕' '좌빨뽕' 등 '뽕쟁이'가 될 판이다. 이 합성어의 사용은 분명히 은유적으로 기술되는 사람을 심하게 비난하고 인격을 폄하할 의도를 담고 있다. 바로 지금이 사용을 자제할 적기가 아닐까.

뉴스 인터뷰 대화 내 ‘네/예’의 기능과 특성 - 인터뷰어 발화를 중심으로

손지연 (연세대학교)

I. 서론

일반적인 대화(ordinary conversation)에서와 달리, 의료, 보건, 응급구조대 등 기관 내에서의 대화(institutional interaction)에서는 기관의 업무(task)와 정체성(identity)에 따른 특정한 목적을 달성하기 위하여 말차레와 행위가 이루어진다(Heritage & Clayman, 2010). 그중 뉴스 인터뷰 대화는 인터뷰어와 인터뷰이 외에 제3의 시청자가 존재하고, 인터뷰어는 대중을 대신해 인터뷰이로부터 정보를 끌어내는 역할을 맡는다(Heritage, 1985; Clayman, 2002; Clayman & Heritage, 2002). 그리고 이 과정에서 인터뷰어에게 중립성(neutrality)이 요구된다는 점에서 토크쇼나 일상적인 대화와 구별된다(Clayman & Heritage, 2002; Heritage & Clayman, 2010). 따라서 시청자들에게 정보를 전달한다는 뉴스 인터뷰의 목적 아래 질문과 대답을 중심으로 말차레 교체(turn-taking)가 일어나며, 이때 인터뷰어와 인터뷰이는 이러한 기관의 정체성에 따라 각각 질문과 대답만을 하도록 제한된다(Greatbatch, 1988). 또한 앞서 언급하였듯 인터뷰어는 전문적인 언론인으로서 중립적인 태도가 요구되기에 인터뷰이의 답변에 대해 개인적인 견해를 내세우는 행동을 피해야 하며(Heritage & Clayman, 2010), 따라서 인터뷰어는 대부분의 말차레를 질문만을 하도록 제한된다(Heritage & Roth, 1995). 또한 일반적으로 영어를 대상으로 한 뉴스 인터뷰의 경우, 인터뷰어와 인터뷰이 각각의 말차레 사이에 ‘uh huh’, ‘yeah’, ‘oh’와 같은 담화표지(receipt token)가 나타나지 않으며, 만약 이러한 담화표지가 나타날 경우 인터뷰이의 말에 동의하는 태도로 해석되어 중립성을 해칠 수 있으므로 부적절한 것으로 취급된다(Heritage & Clayman, 2010).

만약 뉴스 인터뷰 대화 참여자들이 자기 말차레가 아닐 때 발언하거나, 용인할 수 있는 질문과 대답 수준을 넘어서는 등 본래 뉴스 인터뷰 대화의 원형에서 벗어날 경우 청자는 이를 부정적으로 대하고, 화자 역시 자신의 행위가 민감하고 문제가 될 수 있음을 알고 있다(Heritage & Clayman, 2010). 그리고 이러한 화자와 청자의 반응은 곧 뉴스 인터뷰 대화가 철저하게 질문-대답 연속체(Q&A sequence)로 이루어지는 것이 일반적이라는 그들의 인식을 대변한다. 하지만 한국의 뉴스 인터뷰 대화에서는 인터뷰어와 인터뷰이 모두가 담화표지 ‘네’와 ‘예’를 대부분의 말차레에서 사용하는 모습을 보인다. 이러한 담화표지의 사용은 언어 문화적 차이에서 비롯될 수도 있으나, 이와 관계없이 인터뷰어의 중립적 태도가 중요하게 여겨지는 뉴스 인터뷰 대화에서 ‘네’, ‘예’와 같이 상대방의 말에 동의하는 태도로 해석될 수 있는 담화표지의 사용은 분명 인터뷰어에게 요구되는 중립적 태도와 모순되는 것처럼 보인다.

하지만 뉴스 인터뷰와 같이 인터뷰어에게 중립적인 태도가 요구되는 제도적 상황 속에서 인터뷰어가 수없이 많은 청자반응을 보이고 있음에도 불구하고 인터뷰이나 시청자들이 이에 대해 부정적인 반응을 표하지 않는다는 점에서, 한국어 뉴스 인터뷰에서 인터뷰어가 사용하는 청자반응 ‘네’와 ‘예’가 일상 대화에서와 달리 받아들여지는 이유에 대해 탐구할 필요가 있다. 따라서 본 논문에서는 뉴스 인터뷰 대화에서 사용된 인터뷰어의 청자반응 ‘네’와 ‘예’의 기능을 상세히 밝히고, ‘네’와 ‘예’의 특성을 바탕으로 이들이 일반적인 청자반응과 달리 인터뷰어의 중립성을 해치지 않는 것으로 받아들여지는 이유에 대해 알아보고자 한다.

2. 데이터와 방법론

본 논문에서는 2021년 4월부터 10월 사이에 진행된 JTBC 뉴스, 인터뷰 자료 6편을 직접 전사하여 분석하였으며, 각 인터뷰에 대한 정보는 아래 <표 1>에 기술하였다. 총 59분 21초 분량에 해당하는 6편의 인터뷰는 JTBC 뉴스룸에서 한국어로 진행되었으며, 편집 없이 실시간으로 녹화된 인터뷰 영상들만을 대상으로 하였다. 수집된 인터뷰는 필자에 의해 직접 전사되었으며, 사용된 전사기호는 부록에 제시하였다.

Table. 1. Length, Participants, and Date of Each Interview

<표 1> 각 인터뷰의 길이, 대상, 날짜

	1	2	3	4	5	6
총 길이	12:37	12:24	11:26	5:38	8:13	9:03
인터뷰어	오대영	오대영	오대영	서복현	서복현	서복현
인터뷰이	이정재	여홍철, 여서정	윤석열	백순영	성백린	권덕철
날짜	2021.10.05	2021.08.09	2021.07.14	2021.05.05	2021.04.23	2021.05.25

본 논문에서는 인터뷰어가 사용한 ‘네/예’의 분포의 통계적 유의미성을 확인하고, 여러 예시를 통해 ‘네/예’의 기능을 중립성과 중의성을 중심으로 설명한다. 우선 6편의 뉴스 인터뷰 대화에서 인터뷰어가 사용한 모든 ‘네/예’의 분포를 예외 없이 살펴보았으며, 그 결과 총 193번의 ‘네/예’가 나타남을 확인하였다. 우선 구조적 측면에서, 그중 8번은 인터뷰 대화의 도입부에, 9번은 인터뷰 대화의 종결부에 나타났으며, 나머지 176번은 인터뷰 대화 도중에 나타났다. 또한 말차례와 관련하여, 총 193번의 ‘네/예’ 중 125번은 인터뷰이가 한 문장을 모두 발화하여 말차례의 교체가 일어날 수 있는 지점에 나타났으며, 나머지 51번은 인터뷰이가 문장을 발화하는 중간에, 즉 인터뷰이의 말차례가 끝나지 않았을 때 나타났다. 그리고 기능적 측면에서, 말차례의 교체가 일어날 수 있는 지점에서 나타난 125번의 ‘네/예’ 중 허용표지(continuer)로 기능한 경우가 61번, 말차례의 교체가 일어난 경우가 64번이었으며, 인터뷰이의 발화 도중에 나타난 51번의 ‘네/예’는 모두 허용표지로서만 기능하였다. 이후 본 논문에서는 인터뷰어의 발화에서 ‘네/예’가 나타난 지점의 전후를 포함하여 발제한 예시를 토대로 ‘네/예’가 중의성을 가지는 경우와 가지지 않는 경우를 ‘네/예’의 중립성과 관련지어 설명하는데, ‘네/예’의 분포가 내용의 중심이 되는 본 논문의 특성상 문맥이 불필요하거나, 문맥 파악에 문제가 없을 때는 중략하였음을 알린다.

3. 분석결과

3.1. ‘네/예’의 기능: 말차례 시작 표지

뉴스 인터뷰에서 인터뷰어의 역할은 앞서 서론에서 언급하였듯 중립적인 태도를 유지하면서도 시청자들이 원하는 정보를 인터뷰이로부터 끌어내는 것이다. 그렇기 때문에 뉴스 인터뷰라는 기관 내의 대화에서 ‘네/예’와 같은 표지를 사용하는 것은 자칫 인터뷰이의 말에 동조하는 것으로 해석되어 중립성을 해칠 수 있기 때문에 미국이나 영국의 뉴스 인터뷰에서는 지양되는 모습을 확인할 수 있다. 반면 한국의 뉴스 인터뷰에서 인터뷰어의 ‘네/예’는 대부분의 말차례에서 일관적으로 사용될 뿐만 아니라 인터뷰이나 시청자들로부터 아무런 비난을 받지 않기 때문에 인터뷰어의 ‘네/예’가 중립성을 해치는 다른 표지들과는 달리 중립성을 지니고 있음을 짐작할 수 있다. 본 논문에서는 이러한 ‘네/예’의 중립성이 인터뷰어의 일관적인 ‘네/예’의 사용과, ‘네/예’가 가지는 중의성에서 비롯된다고 분석한다.

문장과 같이 하나의 문법 단위(grammatical unit)가 끝나거나 화자가 말을 멈추는 지점은 교체 적정 지점(transition relevance place)으로, 다른 참여자로서의 말차례 교체가 일어날 수 있는 지점이다(Sacks et al., 1974). 그러나 일반적인 대화에서는 모든 교체 적정 지점에서 말차례의 교체가 일어날 수 있지만, 뉴스 인터뷰 대화에서는 인터뷰어의 질문이 끝나기 이전과 인터뷰이의 대답이 끝나기 이전에 많은 교체 적정 지점이 있음에도 불구하고 말차례를 취하거나 *uh huh, yeah, ok* 등의 담화 표지를 사용하지도 않는다(Heritage & Clayman, 2010). 하지만 본 논문에서 다루는 6편의 뉴스 인터뷰 대화에서 나타난 78개의 질문-대답 연속체 중 ‘네/예’와 함께 질문-대답 연속체가 시작된 경우는 약 82%에 해당하는 64번으로, ‘네/예’와 함께 질문-대답 연속체가 시작되지 않은 경우보다 유의미하게 많았다($p = .005$).

Paired Samples T-Test					
Measure 1	Measure 2	t	df	p	Cohen's d
네	- 그 외	4.675	5	0.005	1.908

Note. Student's t-test.

Figure 1. Comparison Between the Questions Initiated by 'ne/ye' and the Others

[그림 1] '네/예'와 함께 질문이 시작된 경우와 그 외의 경우

일반적으로 도입부와 종결부를 제외한 나머지가 질문-대답 연속체로만 이루어지는 인터뷰 대화에서 대부분의 질문-대답 연속체가 인터뷰어의 '네/예'와 함께 시작하였다는 점, 그리고 이러한 '네/예'가 일반적으로 인터뷰어의 중립성을 해치기 때문에 사용이 지양된다는 점에서, 한국어 뉴스 인터뷰에서 인터뷰어가 사용하는 '네/예'가 일반적인 뉴스 인터뷰에서와 달리 중립성을 해치는 표지로 받아들여지지 않는 이유를 탐구해야 할 필요가 있다.

우선 각 인터뷰 대화의 도입부에서 인터뷰어의 '네'가 어떠한 양상을 보이는지 살펴보겠다. 다음은 오대영 앵커가 배우 이정재와 진행하였던 인터뷰 대화의 도입부와 첫 번째 질문이다.

(1) JTBC 뉴스룸, 2021/10/05 (IR: 오대영 앵커, IE: 이정재 배우)

01 IR: .hh 네 돌풍이라는 표현으로도 부족한 한국: 드라마입니다. 오징어

02 게임:..hh 굳이 심각하게 분석하지 않아도 그 느낌대로만 보면 마지막엔:

[중략]

08 IE: =허허

09 IR: <사백오십육>번 들으면 지금도 깜짝깜짝 놀라실 것 같은데요?

대화 예시 (1)의 1번 발화가 이루어지는 도입부에서 인터뷰어는 아직 대화에 참여하지 않은 상태이다. 따라서 해당 시점에서는 인터뷰어만이 발언권을 가지고 있기에 1번 발화에서 사용된 인터뷰어의 '네'는 허용표지가 아니며, '네'를 사용하지 않고도 자신의 말차례를 시작할 수 있다. 실제로 9번 발화에서처럼, 도입부가 종료되고 첫 번째 질문이 시작되는 것이 당연한 상황에서 인터뷰어는 드물지만 '네'를 사용하지 않고 말차례를 시작하기도 한다. 그런데도 1번 발화에서 인터뷰어가 '네'와 함께 도입부를 시작하는 것은 '네'와 말차례의 시작 간에 관련성이 있음을 시사한다.

(2) JTBC 뉴스룸, 2021/08/09 (IR: 오대영 앵커, IE1: 여홍철 선수)

01 IR: 네 지금 스튜디오에 아주: (.) 특별한 (.) 두 분이 (.) 나와

02 있습니다.

03 (1.0)

04 한 명은: (0.5) 도마의 신이고 (0.5) 또 한 명은 (.) 원조

[중략]

13 IE1: =네

14 (.)

15 IR: 자: 두 분이 이렇게 함께: 뉴스 스튜디오에 나오신 건

16 (0.5)

17 처음이죠=

한편 대화 예시 (2)은 오대영 앵커가 운동선수 여홍철, 여서정과 진행하였던 인터뷰 대화의 도입부와 첫 번째 질문이다. 1번 발화에서 인터뷰어는 '네'와 함께 도입부를 시작한다. 앞서 살펴본 대화 예시 (1)와 마찬가지로 1번 발화에서 사용된 '네'는 허용표지가 아니며, 인터뷰어는 아직 대화에 참여하기 전이기 때문에 인터뷰어가 청자반응인 '네'를 사용할 필요가 없음에도 불구하고 인터뷰어는 '네'와 함께 자신의 말차례를 시작한다. 또한 첫 번째 질문이 시작되는 15번 발화에서 인터뷰어는 '네' 대신 '자'와 함께 질문을 시작하는데, 이를 통해서도 '네'의 기능을 유추해볼 수 있다. 김영철(2008), 신철범(2011), 강소영(2014) 등에 따르면 '자'는 화제의 시작을 나타내는 기능을 하는데, 다른 인터뷰들에서는 인터뷰어가 '네'와 함께 첫 번째 질문을 시작하는 모습을 보였다. 이는 '네'가 '자'와 상보적 분포를 이루는 것이 아님을 뜻하며, 곧 '네'가 '자'를 대체할 수 있다는 의미이다. 따라서 해당 위치에서 사용된 '자'는 '네'가 말차례의 시작을 알리는 표지로 기능함을 보여주는 또 다른 증거로 볼 수 있다.

3.2. ‘네/예’의 기능: 허용 표지

하지만 인터뷰이가 발화한 문장이 끝난 뒤 나타나는 모든 ‘네/예’가 인터뷰어의 말차례, 즉 질문의 시작으로 이어지는 것은 아니었다. 이 경우 인터뷰어의 ‘네/예’가 허용표지로 받아들여졌기 때문이라고 해석할 수 있다. 허용표지는 “청자가 화자의 이야기를 듣고 있으며, 화자가 이야기의 주도권을 유지하면서 발화를 계속할 것을 바란다는 신호”인데(손현정 2013), 일반적으로 허용표지는 하나의 단위 경계(unit boundary) 뒤에서 나타난다.

실제로 인터뷰이의 말차례 도중 나타난 인터뷰어의 ‘네/예’의 경우 반드시 허용 표지로 기능하였다. 다음은 서복현 앵커가 성백린 백신 실용화 기술개발 사업단 단장과 진행하였던 뉴스 인터뷰 중 일부이다.

(3) JTBC 뉴스룸, 2021/04/23 (IR:서복현 앵커, IE:성백린 단장)

- 73 IE: 그렇게 되면은 이것은 윤리적인 문제도 (.) 발생이 될 수가 **있고**
74 IR: **예:=**
75 IE: =.hh 그래서 이것은 아: 아: 임상을 진행하는?, (0.5) .h 그 나라와
76 (0.5)
77 아: 정치적인 이 레벨에서 (.) 어 이런 협상도 필요한 (.) 다각적인
78 이와 같은 공조가. (0.5) 필요할 **것이고**,
79 IR: **[예:]**
80 IE: [이런 부분들이 (.) .h 완벽히 이뤄지며는 **가능하겠지만?**,
81 IR: **예:=**
82 IE: =이런 부분들은 (.) 아직 우리나라에서 특별한 경험들이 없기 **때문에**
83 IR: **예:**
84 IE: 이: 이: 앞으로 (.) 같이 숙- 풀어야 될 숙제가 아닌가 생각을
85 합니다?, .hhh

대화 예시 (3)의 73번~85번 발화까지 인터뷰이의 답변이 이어지고 있는데, 73번 발화는 “있고”로 끝나며 아직 문장이 완결되지 않은 상태이다. 이어서 74번 발화에서 인터뷰어는 “예”라고 말하는데, 이때 인터뷰이는 아직 답변을 끝마치지 않은 상황이기 때문에 인터뷰어의 “예”는 허용표지로만 기능한다. 78번 발화의 “것이고”, 80번 발화의 “가능하겠지만”, 82번 발화의 “때문에” 역시 아직 문장이 완결되지 않은 상태이기 때문에 이후 나타나는 인터뷰어의 “예”는 언제나 허용표지로만 기능하고 있음을 확인할 수 있다. 이때 “-고”, “-지만”, “때문에” 등은 두 문장을 이어주는 연결 어미이기 때문에, 연결어미로 이어진 두 문장은 전체적으로 보면 하나의 문장이지만 실제로는 두 부분으로 나뉘기에 이 역시 일반적인 단위 경계에 해당하고, 따라서 인터뷰어의 ‘네/예’가 허용표지로 기능할 수 있다.

이때 인터뷰어의 ‘네/예’는 중립성을 해치지 않는 것으로 받아들여지는데, 이는 질문-대화 연속체가 주를 이루는 뉴스 인터뷰 대화의 특성에서 비롯한다. 앞서 살펴보았듯 인터뷰어는 질문만을, 인터뷰이는 대답만을 하도록 제한된 뉴스 인터뷰 상황에서 인터뷰이의 답변이 끝나기 전에 인터뷰어가 자신의 말차례를 시작하기 위해 ‘네/예’를 사용한다면 일반적인 뉴스 인터뷰 대화에서 벗어나는 행위를 한 것으로 간주하고, 부정적인 인상을 주게 된다. 따라서 뉴스 인터뷰라는 특수한 상황에서 인터뷰이의 답변이 끝나기 전 인터뷰어가 사용한 ‘네/예’는 허용표지로 받아들여지고, 따라서 중립성을 해치지 않는다고 분석할 수 있다.

하지만 앞서 살펴보았듯 허용표지가 나타날 수 있는 단위 경계는 말차례의 교체가 일어날 수 있는 지점이기도 하다. 따라서 인터뷰어가 사용하는 ‘네/예’는 허용표지의 기능과 말차례의 시작을 알리는 표지의 기능 모두로 해석될 수 있기 때문에 중의적이고, 따라서 인터뷰어는 이를 적절히 활용하여 중립적 태도를 유지하며 말차례를 주고받는다.

3.3. ‘네/예’의 중의성

본 논문에서 다루고 있는 6편의 뉴스 인터뷰 대화에서 ‘네/예’가 인터뷰이가 발화한 문장이 끝난 뒤 등장하는 경우는 총 125번이었는데, 그중 약 51%에 해당하는 64번은 인터뷰어의 말

차례로 이어졌지만, 나머지 49%에 해당하는 61번은 인터뷰어의 말차례로 이어지지 않았다. 그렇다면 동일하게 문장 끝에 나타난 '네/예'가 어떤 경우에는 말차례로 이어지고, 어떤 경우에는 말차례로 이어지지 않는 이유가 무엇인지 알아볼 필요가 있다.

이는 인터뷰어의 '네/예' 전후에 등장하는 몇 가지 표지들을 매개로 이루어지는 인터뷰어와 인터뷰이 간의 상호작용을 통해 설명될 수 있다. 인터뷰어의 '네/예' 전후에 나타나는 표지에는 말침(pause), 들숨(inhalation)이 있었으며, 인터뷰어의 '네/예' 직후 인터뷰이의 '네'가 이어지는 경우, 그리고 아무 표지도 나타나지 않는 경우 역시 존재했다. 인터뷰어와 인터뷰이는 '네/예' 전후에 나타나는 이러한 표지들을 토대로 상호작용함으로써 말차례를 주고받는 모습을 보인다. 그리고 이러한 상호작용이 가능한 것은 앞서 살펴보았듯 인터뷰어의 '네/예'가 가지는 기능이 중의적이기 때문이라고 분석할 수 있다.

우선 인터뷰어의 '네/예' 전후에 말침 혹은 들숨이 나타나는 경우 인터뷰어의 말차례로 이어진다. 다음은 서복현 앵커가 성백린 백신 실용화 기술개발 사업단 단장과 진행하였던 뉴스 인터뷰 중 일부이다.

(4) JTBC 뉴스룸, 2021/04/23 (IR:서복현 앵커, IE:성백린 단장)

- 17 IR: 예: .hh (0.5) 방금 전에 (.) <정세균> 전 총리 >그리고 최기영<
18 장관의 발언을 (0.5) 같이 들으셨는데요?, (0.5) 바로 좀 질문을
19 드리죠?, 올해 안에 개발이 가능합니까,
20 (0.5)
21 IE: .hhhhh 제가 볼 적에 짐: 어: 장관님과 총리님의 (.) 예: 말씀은
22 (0.5) 우리의 (.) 강력한 의지표명으로 **생각됩니다**,
23 IR: **예**,
24 IE: 사실은 우리의 의지도 중요하지마는?, (.) 환경적인 요인과 (0.5)
[중략]
33 (.) 이런 생각을 **해봅니까?**,
34 (0.5)
35 IR: **네**,
36 (1.5)
37 보통 >왜냐면< 지금 임상:: 임상과 이상은 진행이 일부 되고 있기
[중략]
45 IR: **올해 안에 개발 완료**가 되기는 어렵다는 말씀이신 건가요?

대화 예시 (4)의 22번 발화에서 인터뷰이는 "생각됩니다"로 문장을 끝마치는데, 문장이 끝나는 지점은 일반적인 단위 경계에 해당하기 때문에 23번 발화에서 인터뷰어의 '예'가 허용표지와 말차례의 시작을 알리는 표지 모두로 기능할 수 있는 지점이다. 하지만 이때 인터뷰어의 '예' 직전에 인터뷰이가 말을 쉬지 않았기 때문에 인터뷰어는 인터뷰이의 답변이 아직 끝나지 않았다고 판단하게 되고, 따라서 들숨을 쉬는 등 발화를 시작하려는 신호를 보내지 않는다. 한편 인터뷰이 역시 아직 자신의 답변을 마치지 않았기 때문에 인터뷰어의 '예'는 허용표지로 받아 들여지고, 인터뷰이는 24번 발화부터 답변을 계속 이어 나간다. 24번~33번 발화까지 이어진 인터뷰이의 답변 역시 "해봅니까"로 끝마치게 되는데, 마찬가지로 해당 위치는 문장이 끝나는 지점으로 일반적인 단위 경계에 해당한다. 따라서 35번 발화에 나타난 인터뷰어의 '네'가 허용표지와 말차례의 시작을 알리는 표지 모두로 기능할 수 있는 지점이다. 하지만 '네' 직전에 인터뷰이가 0.5초간 말을 쉬고 있고, 인터뷰어의 '네' 이후에도 1.5초간 말을 쉬는 모습을 보인다. 인터뷰어는 인터뷰이의 대답이 완전히 끝나 자신의 '네'를 말차례의 시작을 알리는 표지로 받아들였다고 판단하여 자신의 말차례를 시작하는 모습을 볼 수 있다.

다음은 오대영 앵커가 배우 이정재와 진행하였던 뉴스 인터뷰 중 일부이다.

(5) JTBC 뉴스룸, 2021/10/05 (IR:오대영 앵커, IE:이정재 배우)

- 26 IR: **[예]**.
27 IE: [((laugh))]
28 (0.5)

29 IR: .hhh 80여 개국에서 지금 1위를 하고 [있다 (.) 라고 하더라고요.
 30 IE: [((laugh))
 31 IR: .hh 어: 왜 이렇게: (.) 전세계적으로 열광을: (.) 할까요?
 [중략]
 55 IR: [아: 이제 시작할 수도 있다?,
 56 IE: [((laugh)) 네:.=
 57 IR: =예. .hhhhh 어:: .hhh (.) 드라마 쪽 보면 (.) 한국의 달고나::
 [중략]
 70 오랫동안 했다면서요?
 71 IE: .hhh 예 모든 게임 장면들을 다 오랫동안 촬영을 하긴 (.) 했는데:
 [중략]
 83 죽:어라 알아야되는 (.) 그런: 이제 연기를=이제 하게됐죠=
 84 IR: =예. .hh 그때 어 촬영하면서 날이 많이 덥고 습해가지고 이게 달고나가
 85 많이 (.) 녹았다면서요?
 86 IE: .hhhhh 예 한여름에 찍게:: (.) 됐었던 (.) 장면이었는데, 이제: 잘
 [중략]
 101 분들이 이제 기억이 나네요.
 102 IR: 예. .hhh 자 인생 벼랑 끝에 서있는 사람들을 모아서 사백: 오십: 육억
 103 원의 상금을 걸고 게임을 하는, (.) 그런 설정입니다. .hhh 게임::을
 [중략]
 108 주제: (.) 의식이 나- 물어나는 장면:=
 109 IE: =네네=
 110 IR: =이 (.) 뭐가 있을까요?

대화 예시 (5)의 26번 발화에서 인터뷰어는 '예'를 사용하는데, 28번 발화에서 0.5초 간의 말 침이 나타나고, 29번 발화에서 들숨을 마시며 자신의 말차례를 시작하는 모습을 보인다. 이는 0.5초간 말을 침으로써 인터뷰이의 답변이 완전히 끝났는지, 그리고 자신의 '예'를 어떠한 표지로 받아들였는지 확인하기 위함이며, 인터뷰이의 답변이 이어지지 않자 인터뷰어는 들숨을 마시며 인터뷰이에게 자신이 이제 발화를 시작할 것이라는 신호를 보낸다. 57번 발화에서도 인터뷰어는 '예' 뒤에 들숨을 마시며 자신의 말차례를 시작하는 모습을 보이며, 84번, 102번 발화에서도 동일한 양상을 보인다.

두 번째로, 인터뷰어의 '네/예' 뒤에 인터뷰이의 '네'가 등장하는 경우 인터뷰어의 말차례로 이어진다. 다음은 오대영 앵커가 윤석열 대통령의 후보자 시절 진행하였던 인터뷰 중 일부이다.

(6) JTBC 뉴스룸, 2021/07/14 (IR:오대영 앵커, IE:윤석열 전 대통령 후보)
 09 IR: 네. 정치 선언 (0.5) 한 지가=이제 (.) 보름됐습니까?
 10 IE: 그- 어 보름됐습니다.
 11 IR: 예;
 12 IE: ['네.]
 13 IR: [.hhh] 어: 검사일 때와 어: 그 이후에 (.) 가장 많이 달라진 게
 14 뭘까요?
 15 (2.0)
 16 IE: 아무래도 새로운 일이나까?, (1.0) 아 제 입장에서 보면은 뭐: 모든 게
 [중략]
 20 IE: 하여튼 뭐 열심히 (.) 하고 있습니다.
 21 IR: 네.
 22 IE: 네.
 23 IR: 어:: 정치선언 이후에
 24 IE: 네,
 25 IR: 지지율 신경이 더 많이 쓰이지요?
 [중략]
 51 IR: [예.
 52 IE: '네
 53 IR: .hh 그 대안이 안 보인다는: <말씀>을 많이

[중략]

60 IR: [이런 얘기도 나오고 있는데?, .hh 그런 비판:도 어떻게 받아
61 들이시는지요?

대화 예시 (6)의 9번 발화에서 인터뷰어의 첫 번째 질문이 끝나고, 11번 발화에서 인터뷰어는 '예'를 사용한다. 이후 12번~13번 발화에서 인터뷰어의 '네'와 인터뷰어의 들숨이 동시에 나타난다. 이때 인터뷰어의 '네'는 추가적인 답변의 일부가 아니라, 자신의 답변이 끝났으며 인터뷰어가 발언권을 가져가도 좋다는 허용표지로서 기능하기 때문에 인터뷰어는 이를 받아들여 들숨을 쉬며 자신의 말차레를 이어가는 모습을 확인할 수 있다. 21번 발화에서도 인터뷰어의 '네' 이후 22번 발화에서 인터뷰어의 허용표지로서의 '네'가 이어지고, 인터뷰어의 말차레가 이어지는 모습을 확인할 수 있다. 51번 발화에서도 인터뷰어의 '예' 이후 52번 발화에서 인터뷰어의 허용표지로서의 '네'가 이어지고, 53번 발화에서 인터뷰어가 들숨을 쉬며 인터뷰이에게 이제 자신이 말차레를 시작할 것임을 보이고 있음을 확인할 수 있다.

따라서 인터뷰어의 '네/예' 전후에 말을 쉬거나 들숨이 나타나는 경우, 그리고 인터뷰어의 '네/예' 이후 인터뷰이의 허용표지 '네'가 이어지는 경우 인터뷰어의 '네/예'가 말차레의 시작을 알리는 표지로 기능하여 자신의 말차레로 이어지며, 이러한 말차레 교체는 '네/예'의 중의적 기능과 인터뷰어와 인터뷰이 간의 끊임없는 상호작용을 통해 가능하다는 것을 확인할 수 있었다. 실제로 6편의 뉴스 인터뷰 대화에서 인터뷰어의 '네/예'가 문장 끝에서 말차레로 이어졌던 64번의 경우 중 인터뷰어의 '네/예' 전후에 말쉴, 들숨이 나타나거나 인터뷰이의 허용표지 '네'가 이어진 경우가 약 94%에 해당하는 60번이었으며, 그렇지 않은 경우보다 유의미하게 많았다($p = .010$).

Paired Samples T-Test					
Measure 1	Measure 2	t	df	p	Cohen's d
말쉴, 들숨 등	- x	4.050	5	0.010	1.653

Note. Student's t-test.

Figure 2. Comparison between 'ne/ye' with Pause or Inhalation and the others

[그림 2] '네/예' 전후에 말쉴과 들숨이 나타난 경우와 그 외의 경우

정리하자면, 인터뷰어는 뉴스 인터뷰의 도입부, 질문-대답 연속체, 종결부 모두에서 '네/예'를 두루 사용하며, 이때 '네/예'는 말차레의 시작을 나타낼 수도 있고, 허용표지로서 기능할 수도 있다는 점에서 중의적이다. 만약 인터뷰어가 매 질문을 시작할 때마다 '네/예' 대신 '자'와 같이 말차레의 시작을 알리는 단일한 기능을 가진 표지를 사용한다면, 인터뷰이로부터 충분한 정보를 끌어낼 수 없을 가능성이 커지고, 인터뷰이의 말차레를 지나치게 통제한다는 점에서 중립적이지 않은, 독선적인 태도로 여겨질 가능성이 커진다. 따라서 인터뷰어는 문장이 끝난 뒤 인터뷰이가 말을 쉬는지, 자신의 '네/예' 뒤에 인터뷰이의 답변이 이어지는 않는지 등 여러 표지를 토대로 자신의 말차레를 시작해나가며, 인터뷰이는 자신의 답변을 통해 전달하고 싶은 정보를 모두 전달했다면 인터뷰어의 '네/예'를 말차레의 시작을 알리는 표지로 받아들이고, 자신이 전달하고 싶은 정보가 아직 남았다면 허용표지로 받아들여 발언을 계속해나간다. 결국 '네/예'는 인터뷰어의 중립적 태도를 해치지 않는 표지로 기능하며 인터뷰어가 정보 전달자라는 자신의 역할에 충실하게 하고, 대화자 간 상호작용이 중요한 기관 내의 대화에서 말차레 교환을 더욱 용이하게 해주는 장치로써 활용되고 있음을 알 수 있다.

4. 결론

본 논문에서는 한국어 뉴스 인터뷰 대화 6편을 분석하여, 뉴스 인터뷰 대화에서 사용된 인터뷰어의 청자반응 '네'와 '예'의 기능을 상세히 밝히고, '네'와 '예'의 특성을 바탕으로 이들이 일반적인 청자반응과 달리 인터뷰어의 중립성을 해치지 않는 것으로 받아들여지는 이유에 대해 알아보았다. 구체적으로, 본 논문에서는 인터뷰어가 사용한 '네/예'의 분포의 통계적 유의미성을 확인하고, '네/예'의 중립성이 인터뷰어의 일관적인 '네/예'의 사용과, '네/예'가 가지

는 기능의 중의성에서 비롯된다고 분석하였다. 그리고 인터뷰어와 인터뷰이가 이를 어떻게 활용하여 상호작용하며 말차례를 교환해나가는지를 여러 예시를 통해 살펴보았다.

인터뷰이가 발화한 문장이 끝난 뒤 나타난 인터뷰어의 ‘네/예’는 우선 말차례의 시작을 나타내는 표지로 기능할 수 있음을 보였다. 인터뷰어는 아직 인터뷰이가 발언권을 가질 수 없는 도입부에서도 자신의 말차례를 ‘네/예’와 함께 시작하는 모습을 보였으며, 대부분의 질문을 ‘네/예’와 함께 시작하는 모습을 보였다. 첫 번째 질문의 경우 역시 인터뷰이가 아직 발언권을 가지기 전이기 때문에 인터뷰어는 아무런 표지 없이 자신의 말차례를 시작하는 것이 가능하지만 드물게 ‘네/예’ 대신 ‘자’를 사용하기도 하는데, 이때 ‘자’는 말차례의 시작을 알리는 기능을 가진다는 점에서 동일한 위치에서 사용된 ‘네/예’ 역시 말차례의 시작을 알리는 기능을 내포하고 있음을 알 수 있었다.

하지만 인터뷰이가 발화한 문장이 끝난 뒤 나타난 모든 ‘네/예’가 인터뷰어의 말차례, 즉 질문-대답 연속체의 시작으로 이어지지는 않았다. 이때 인터뷰어의 ‘네/예’는 허용표지로서 받아들여졌는데, 인터뷰어의 ‘네/예’ 전후에 나타나는 말썽, 들숨 등의 표지와 인터뷰어의 ‘네/예’ 직후 이어지는 인터뷰이의 ‘네’ 등을 통해 인터뷰어와 인터뷰이가 상호작용하며 말차례를 교환해나가기 때문임을 보였다. 인터뷰어의 ‘네/예’ 전후에 아무런 표지가 나타나지 않을 경우 인터뷰어는 인터뷰이의 답변이 끝나지 않았다고 판단해 자신의 말차례를 시작하지 않고, 인터뷰이 역시 자신의 답변이 끝나지 않았기 때문에 이를 허용표지로 받아들인다. 반면 인터뷰어의 ‘네/예’ 전후에 말썽이나 들숨이 나타나는 경우, 인터뷰어는 인터뷰이의 답변이 끝났다고 판단해 자신의 말차례를 이어가는 모습을 보였다. 인터뷰어의 ‘네/예’ 직후 인터뷰이의 ‘네’가 이어지는 경우 이때 인터뷰이의 ‘네’는 자신의 답변이 끝났으며 인터뷰어가 발언권을 취하여도 좋다는 허용표지로서 기능하며, 인터뷰어 역시 이를 토대로 인터뷰이의 답변이 끝났다고 판단하여 자신의 말차례를 시작하는 모습을 보였다.

본 논문에서 다루었던 6편의 뉴스 인터뷰 대화에서 ‘네/예’와 함께 질문-대답 연속체가 시작된 경우가 약 82%로 그렇지 않은 경우보다 유의미하게 많았으며, 인터뷰어는 인터뷰의 도입부, 질문-대답 연속체, 종결부 모두에서 ‘네/예’를 일관적으로 사용하는 모습을 보였다. 그리고 이때 인터뷰어의 ‘네/예’는 말차례의 시작을 알리는 기능과 허용표지의 기능을 모두 가지기 때문에 중의적이다. 인터뷰어와 인터뷰이는 이러한 ‘네/예’의 중의성을 토대로 서로 끊임 없이 상호작용하며 말차례를 구성해나갈 수 있게 되며, 그 결과 인터뷰어는 중립적인 태도를 잃지 않으면서도 인터뷰이로부터 충분한 정보를 끌어내 청중에게 전달해야 하는 자신의 역할에 충실할 수 있게 된다.

인공지능 언어 모델의 절 연결 능력에 대한 일고찰

-이유를 나타내는 ‘-어서’와 ‘-니까’를 중심으로-

이지은·황동진·송상현 (고려대학교)

1. 연구 주제

- ‘이유’를 나타내는 ‘-어서’와 ‘-니까’

(1) ㄱ. 감기에 걸려서 학교에 가지 못했다.
 ㄴ. 비가 오니까 우산을 가지고 가라.

(2) ㄱ. *일어났어서 샤워를 했다.
 ㄴ. *비가 와서 우산을 가져가라.
 ㄷ. *실은 개인적이 사정이 생겼으니까 휴학을 신청했거든.
 ㄹ. *폭우가 내리니까 늦었습니다.
 ㅁ. *다리가 부러져서 두통약을 먹었다.

- 연구 질문

- 1) 인공지능 언어 모델이 ‘이유’를 나타내는 ‘-어서’와 ‘-니까’에 이어 후행절을 적절히 생성하는가?
- 2) 인공지능 언어 모델이 생성한 후행절이 ‘-어서’와 ‘-니까’에 대해 새롭게 조명해 주는 바가 있는가?

2. 절 연결 분석 시 고려할 사항

- ‘-었-’의 유무
- 절 연결의 세 층위
- 정보관할권
- 담화의 성격
- 상식 혹은 세계지식

3.연구 방법

- ChatGPT:

ㄱ. gpt-3.5-turbo-0301(터보모델)

ㄴ. text-davinci-003(다빈치모델)

- Temperature: 0.5

- One-shot

(3) 프롬프트: 제시문 뒤에 적절하게 이어질 내용을 이야기해 주세요.

제시문 예시: 홍수가 나면서

생성 결과 예시: 홍수가 나면서 다리가 물에 잠겼다.

제시문: {original_prompt}

(4) 시험이라서 열심히 공부해야 했다.

4. 연구 결과

<표1> 터보모델이 생성한 문장의 분류

선행절	후행절			합계
	목표 구문	다중 연결	기타 구문	
-어서	77	20	3	100
-었어서	83	17	0	100
-니까	69	31	0	100
-었으니까	77	23	0	100
합계	306	90	3	400

<표2> 다빈치모델이 생성한 문장의 분류

선행절	후행절			합계
	목표 구문	다중 연결	기타 구문	
-어서	91	9	0	100
-었어서	91	9	0	100
-니까	88	9	3	100
-었으니까	93	6	1	100
합계	363	33	4	400

<표3> 터보모델의 절 연결의 적절성 판단

선행절	절 연결의 적절성		합계
	적절	부적절	
-어서	56(72.73)	21(27.27)	77(100)
-었어서	57(68.67)	26(31.33)	83(100)
-니까	48(69.57)	21(30.43)	69(100)
-었으니까	21(27.27)	56(72.73)	77(100)
합계	182(59.48)	124(40.52)	306(100)

<표4> 다빈치모델의 절 연결의 적절성 판단

선행절	절 연결의 적절성		합계
	적절	부적절	
-어서	59(64.84)	32(35.16)	91(100)
-었어서	60(65.22)	31(34.07)	91(100)
-니까	45(51.14)	43(48.86)	88(100)
-었으니까	42(45.16)	51(54.84)	93(100)
합계	206(56.75)	157(43.25)	363(100)

<표5> 터보모델의 오류문 양상

오류의 유형		연결어미			
		-어서	-었어서	-니까	-었으니 까
의미	이유	6	6	12	11
	상식	4	5	1	3
	총합	10	11	13	14
		(47.61)	(42.3)	(61.9)	(25)
표현	구문	1	3	3	5
	명사	-	1	1	-
	서술어	1	2	-	1
	총합	2	6	4	6
		(9.52)	(23.08)	(19.05)	(10.71)
합계		(38.71)	(14.52)		

문법	문장부사	-	-	-	1	1
	부정극어	-	-	-	1	1
	시제	-	-	1	19	20
	연결어미	-	-	-	9	9
	인칭	-	2	-	-	2
	접속사	-	1	-	-	1
	조사	-	-	-	1	1
	피동	-	1	-	-	1
다중	총합	0	4	1	31	36
		(0)	(15.38)	(4.76)	(55.36)	(29.03)
	이유&서술어	2	-	1	-	3
	이유&조사	1	-	-	-	1
	상식&수식어	1	-	-	-	1
	구문&시제	-	-	-	1	1
	서술어&수식어	1	-	-	-	1
	조사&명사화소	1	-	-	-	1
이해불가	총합	6	0	1	1	8
		(28.57)	(0)	(4.76)	(1.79)	(6.45)
		3	5	2	4	14
합계		(14.29)	(19.23)	(9.52)	(7.14)	(11.29)
		21	26	21	56	124
		(100)	(100)	(100)	(100)	(100)

<표6> 다빈치모델의 오류문 양상

오류의 유형		연결어미			
		-어서	-었어서	-니까	-었으니 까
의미	상식	3	2	5	2
	이유	9	15	17	13
	총합	12	17	22	15
		(37.5)	(54.84)	(51.16)	(29.51)
표현	구문	-	-	5	3
	명사	2	-	-	1
	서술어	3	4	1	-
	총합	5	4	6	4
		(15.62)	(12.9)	(13.95)	(7.84)
문법	높임	-	-	3	-
	서술어형태	-	1	1	2
	대명사	1	-	-	1
	조사	3	1	-	2
	성분생략	-	2	-	-
	시제	-	-	3	12
	연결어미	-	-	2	6
	인칭	-	-	-	1
	피동	2	-	1	-
	총합	6	4	10	24
		(18.75)	(12.9)	(23.26)	(47.06)

다중	상식&구문	-	1	-	-	1
	상식&양태	-	-	1	-	1
	이유&서술어형태	-	-	-	1	1
	이유&서술어	-	-	-	1	1
	이유&조사	-	1	-	1	2
	구문&연결어미	-	-	1	-	1
	높임&서술어형태	1	-	-	-	1
	서술어&연결어미	-	-	-	1	-
	시제&부정문	-	-	-	1	1
	시제&서술어	-	-	-	1	1
	시제&조사	-	1	-	-	-
	총합	1	3	2	6	12
		(3.13)	(9.68)	(4.65)	(11.76)	(7.64)
이해불가		8	3	3	2	16
		(25)	(9.68)	(6.98)	(3.92)	(10.19)
합계		32	31	43	51	157
		(100)	(100)	(100)	(100)	(100)

4.1 문장의 유형

<표7> 터보모델의 문장의 유형에 따른 분류 (적절)

선행절	문장 유형				총합
	평서문	의문문	명령문	청유문	
-어서	56	-	-	-	56
-었어서	57	-	-	-	57
-니까	48	-	-	-	48
-었으니까	21	-	-	-	21
합계	182	0	0	0	182

<표8>인지문장의 유형에 따른 분류 (적절)

선행절	문장 유형				총합
	평서문	의문문	명령문	청유문	
-어서	59	-	-	-	59
-었어서	60	-	-	-	60
-니까	44	-	1	-	45
-었으니까	41	-	1	-	42
합계	204	0	2	0	206

4.2 후행절의 시제

<표11> 터보모델의 후행절의 시제에 따른 분류 (적절)

선행절	후행절의 '-었' 유무		합계
	없음	있음	
-어서	3(5.36)	53(94.64)	56(100)
-었어서	0(0)	57(100)	57(100)
-니까	10(20.83)	38(79.16)	48(100)
-었으니까	5(23.8)	16(76.19)	21(100)
합계	18(9.89)	164(90.11)	182(100)

<표12> 다빈치모델의 후행절의 시제에 따른 분류 (적절)

선행절	후행절의 '-었' 유무		합계
	없음	있음	
-어서	32(54.23)	27(45.76)	59(100)
-었어서	4(6.67)	56(93.33)	60(100)
-니까	31(68.89)	14(31.11)	45(100)
-었으니까	16(38.1)	26(61.9)	42(100)
합계	83(40.29)	123(59.71)	206(100)

4.3 후행절의 양태표현

<표13> 터보모델의 후행절의 양태에 따른 분류 (적절)

선행절	후행절의 양태 의미								합계 양태/적절 (백분율)
	가 능	불 가 능	추 측	의 도	소 망	결 정	약 속	의 무	
	6	4	-	1	-	1	-	-	12/56
-어서									(21.43)
-었어서	1	5	-	1	-	1	-	3	11/57
									(19.3)
-니까	-	2	4	-	2	-	-	2	10/48
									(20.83)
-었으니까	-	1	2	-	1	-	-	3	7/21
									(33.33)
합계	7	12	6	2	3	2	0	8	40/182
									(21.98)

<표14> 다빈치모델의 후행절의 양태에 따른 분류 (적절)

선행절	후행절의 양태 의미								합계 양태/적절 (백분율)
	가 능	불 가 능	추 측	의 도	소 망	결 정	약 속	의 무	
-어서	2	5	-	2	2	-	-	3	14/59
									(23.73)
-었어서	1	7	-	-	6	-	-	1	15/60
									(25)
-니까	1	3	-	7	6	-	-	2	19/45
									(42.22)
-었으니까		4	2	2	1	2	1	7	19/42
									(45.24)
합계	4	19	2	11	15	2	1	13	67/206
									(32.50)

5. 논의

- (5) ㄱ. ‘이유/원인-추론/결과’의 포착
ㄴ. 서술어의 활용형 및 결합 형태 포착
ㄷ. 동음이의어 및 유의어의 처리
ㄹ. 특정 단어의 처리
ㅁ. 인칭 제약의 포착
ㅂ. 영어 표현의 번역

6. 결론

- 터보모델과 다빈치모델은 후행절을 각각 59.48%와 56.75%로 적절히 생성함.
- 두 모델 모두에서 선행절이 ‘-었으니까’일 때 후행절의 적절성이 가장 낮았음.
- 두 모델 모두에서 문장 유형은 평서문이 압도적으로 많았음.
- 후행절의 양태표현은 ‘불가능’, ‘소망’, ‘의무’, ‘의도’의 의미가 많이 관찰되었음.
- 선행절 속 서술어에 ‘-었-’이 쓰인 경우에 후행절 속 서술어에 ‘-었-’이 나타나는 빈도는 터보모델과 다빈치모델이 연결어미에 따라 서로 다른 양상을 보였음.

아학편(兒學編) <지석영본>(1908)의 구성과 전사·번역 양상

유 성 현 (광운대학교)

1. 『아학편』의 저자와 구성

본고는 『아학편(兒學編)』 <지석영본>(1908)에서의 전사(transcription)와 번역 양상에 대해 탐구한다. 『아학편』은 본디 조선 후기 유학자 정약용(丁若鏞, 1762 - 1836)이 1801년부터 1818년까지 전라남도 강진으로 유배를 떠난 시기에 저술한 한자 학습서이다. 발간 일자는 불명이지만, 정약용의 현손 정규영(丁奎英, 1872 - 1927)은 정약용의 행적을 연대순으로 정리하여 편찬한 책인 『사암선생연보(俟菴先生年譜)』(1921)에서 『아학편』이 1804년(순조 4년) 봄에 완성했다고 주장했다.

(1) 純祖四甲子 公四十三歲春 兒學編訓義成 (凡二千文 - 원주)

순조 4년 갑자년, 정약용이 43세가 되던 해 봄에 『아학편훈의』를 완성하였으니, 모두 합해 2천 자였다.

『아학편』은 『천자문』보다 아동에게 적합한 수준의 한자 학습서의 필요성을 느껴 저술한 책이며, 상권과 하권으로 나뉘어 총 2,000자가 수록되어 있어서 쌍천자문(雙千字文)이라고도 불린다. 여기서 상권에는 구체적인 형상을 지칭하는 유형자(有形字) 1,000자가 수록되어 있으며, 하권에는 형상이 아닌, 상태나 동작을 지칭하는 무형자(無形字) 1,000자가 수록되어 있다.

유형자를 무형자보다 먼저 오게 하는 구성은 정약용의 학습관에서 비롯하였다. 정약용은 아동의 인지능력은 실재하는 사물에 대한 인식에 의해 향상된다고 믿으며, 학습 순서를 고려하지 않고 추상적인 개념을 먼저 제시하는 『천자문』을 학습서로 사용하기를 반대하였다. 사실 정약용 이전에도 많은 학자들이 아동학습서로서의 『천자문』을 비판했는데, 대표적으로 조선 초중기 어문학자 최세진(崔世珍, 1463 - 1542)은 『훈몽자회(訓蒙字會)』(1527)에서 “어린아이가 글을 배워 글자를 알게 하려면 모름지기 먼저 사물과 그 사물을 가리키고 묶어주는 말의 글자를 적어 알게 하”여야 한다 주장한 바가 있다. (이준환, 2014a:33.)

또한 『아학편』의 구성에 영향을 준 한자 학습서로는 『신증유합(新增類合)』(1576)이 꼽히는데, 1책 2권으로 이루어져 있으며, 상권에 사물들을 수록하고 하권에 동작을 나타내는 자들을 기재했다는 점에서 두 책이 동일한 구성이기 때문이다. 『아학편』은 『훈몽자회』나 『신증유합』보다 글자를 적게 수록하는 대신, ㉠ 54개의 많은 하위 부문으로 글자들을 세세히 나눴고, ㉡ 『천자문』과 동일하게 여덟 자마다 운(韻)을 바꾸어 글자들을 외우기 용이하며, ㉢ 서로 대립 또는 호응하는 뜻의 두 자를 단어쌍으로 묶는 구성을 택하였다. 특히 하권의 형용사들을 대립쌍(五十.1580 - 五十四.1723)과 호응쌍(五十四.1724 - 五十七.1812)으로 명료하게 나눈 것이 특징이다.

구성의 이점들에 힘입어, 지석영(池錫永, 1855 - 1935)과 민병석(閔丙奭, 1858 - 1940)은 『아학편』을 어린이를 위한 어학 입문서로 채택해야 한다고 주장했으며, 그 사유를 각각 1905년(광무 9년) 1906년(광무 10년)에 달은 두 서문에서 밝혔다. (띄어쓰기와 번역은 필자에 의함, 이하 동일.)

(2) 凡二千 分有形無形於 人世之日用者迨無所遺 洵童穉入學教科之津筏

(정다산 선생의 책은) 모두 이전 자인데, 유형과 무형으로 구분하면서 우리 삶에서 날마다 쓰는 것들을 거의 모두 다루기까지 하니, 참으로 학문에 발 들이는 어린이에 있어 나뭇배와 같은 교과이다.

(3) 周興嗣千字文 原非爲牖蒙而作者入于 吾韓爲夢學之初讀 是誰之倡也若

주흥사의 『천자문』은 본디 어린이들이 깨치기 위해 쓰인 책이 아니거늘, 우리 한국에 들어와서는 어린이들이 공부하는 데에 있어 처음으로 읽는 책이 되었으니, 이는 누가 주장한 것이란 말인가?

이후 지석영은 1908년(융희 2년)에 용산에서 『아학편』의 새로운 석판본을 출간했으며, 이를 통칭

1) 여기에서는 이준환(2014a)의 기준을 따르고자 한다.




<지식영본>이라 한다. <지식영본>은 단권 63장으로 구성되어 있으며, 서명인 ‘아학편’에까지 주가 달려있으므로 수록자는 2,002자로 늘었다.²⁾ <지식영본>의 가장 큰 변화는 지식영이 외국어에 능통했던 전용규(田龍圭, ? - ?)의 도움을 받아 모든 한자 표제어에 당대의 영어, 중국어, 일본어로 번역한 주(註)를 달았다는 점이다. 지식영은 서문에서 표제어를 4개국어로 표기한 이유를 보기 (4)와 같이 설명하며, 이는 개화기 외국어 교육 광풍을 염두에 둔 것으로 해석된다.³⁾

(4) 此書 釋中西及 東洋音義 使國人在 稗學有所方向

이 책을 해석하여 중국(중국어), 서양(영어), 그리고 동양(일본어)의 음과 뜻으로 풀이하면 우리나라 사람들이 어린이들을 가르치는 데에 방향을 둘 수 있다.

사실 이전 『국한회어(國韓會語)』(1895)에서도 한국어-한자-영어를 서로 대응하는 시도를 보였으나, 『국한회어』는 한한사전(韓漢事典)의 용도로 저술되었으며, 영단어를 구미 지명에 한하여만 붙였다는 점에서 그 방향성이 달랐다. 그러므로 <지식영본>은 『한영즈언(韓英字典)』(1890)을 계승하는 한국의 두 번째 한영 단어집이자 한국 최초의 4개국어 사전이라는 큰 의의를 지닌다.

	① 한국어 훈음	② 중국어 음	③ 중국어 성조	④ 한자
	⑤ 소전자	⑥ 한국어 이표기	⑦ 반절(反切)	⑧⑨ 일본어 훈과 한국어 전사 표기
	⑩⑪ 일본어 음과 한국어 전사 표기	⑫ 영어 번역어에 대한 로마자 표기	⑬ 영어에 대한 한글 전사 표기	

<지식영본>의 표제어에는 상기와 같이 총 4개 국어, 13 종류의 주가 붙으며, 각 쪽마다 16자가 수록되어 있다. 본고에서는 이 가운데 한자로 표기된 부분들(‘④ 한자’, ‘⑤ 소전자’, ‘⑦ 반절’)을 제외한 다국어의 전사와 번역 양상을 개략적으로 살펴본다.

2) 2,003자 중에서 學(‘아학편’의 ‘학’)은 하권에 이미 수록되어 있으므로 중복 수록이다. (一.2, 四十.1255)

3) “영어는 역사적으로 19세기 중후반 서세동점의 문화 충격 속에서 한국이 동아시아의 여러 나라와 함께 겪은 근대적 자의식의 일환이었으며 영어교육에 대한 목마름과 뜨거움은 외국어에 관한 교과 교육의 차원을 훌쩍 뛰어넘는 사회적 현상이었다.” (황영순, 2014:210.)

2. 한국어 훈음 표기

<지석영본>에서의 한국어 표기는 ‘① 한국어 훈음’과 ‘⑥ 한국음 이표기’로 나뉜다. 이 가운데 ‘① 한국어 훈음’에 관하여, 본서 초두에 실린 「대한국문(大韓國文)」이 표기의 기준이 된다. 「대한국문」에서는 ‘새로이 바로잡은 명사 연음에 대해 발음(新訂名詞聯音辨)’이라는 이름으로 새로운 표기법을 제기하는데, 그 일부를 발췌하자면 보기 (5)와 같다.

- (5) 凡做 語之曳聲에 亦加一點 (….) 做語曳聲標 簾 발' 럽 足 발 족 列 버'릴 렬 捐 버'릴 연 之類
배쫓 땃들等 名은 ㅅ字를 中間에 置하여 上下의 名詞로 하야곰 障礙됨이 업'시 聯讀하면 其音義가
了然하니 배ㅅ돏 매ㅅ돌 之類

보기 (5)를 요약하자면, 아포스트로피 '는 소리를 끄는 표지로, 한글 ‘ㅅ’은 합성어 형태소들의 원형을 유지하고 경음화를 방지하기 위한 형태소 분리 표지로 쓰인다는 뜻이다. 특히 본서에서의 ‘ㅅ’ 용법은 이례적이라 할 수 있는데, 합성어 구성 요소로서 중성에 붙거나 ㅅ계 합용병서로서 초성에 결합하는 대신에, 형태소 중간에 ‘ㅅ’이 단독으로 표기되기 때문이다.

- (6) 度 - 찬'ㅅ장 기' (二十三.708) (cf. ‘찬장’ 『신자전』 (1915))
(7) 象 - 코ㅅ길이 상' (十七.522) (cf. ‘코기리’ 『한영조년』 (1890))
(8) 浮 - 뜰ㅅ 부 (五十二.1636) 汎 - 뜰ㅅ 범 (五十六.1704)

‘ㅅ’은 수록자 가운데 39자(표제어의 1.95%)에만 쓰이나, 사실상 관형격 조사 ‘-의’와 비슷한 쓰임을 한다. 보기 (6)에서 (8)은 ‘ㅅ’의 쓰임에 대한 구분이다. 본서에서 ‘ㅅ’은 보기 (6)과 같이 한자 사이에 배치되기도 하며, 보기 (7)과 같이 당대의 표기에 역행하는 과다분절 표기를 초래하기도 한다.⁴⁾ 보기 (8)은 이례로, ‘ㅅ’이 형태소 사이가 아닌 어말에 위치하여 뒤에 붙을 자를 미리 수식한다. 이러한 경우는 보기 (8)에 기재된 두 자에서만 나타난다.

- (9) 主쥬 (二.46) 北북 (三十四.1071) 摘덕 (四十二.1341)
(9') 兒으 (一.1) 牲싱 (二十. 615) 祀스 (四十二.1335)
(10) 灑새 (四十五.1420) 壤괴 (五十六.1762) 拓탁 (五十七.1801)

또 다른 한국어 주해인 ‘⑥ 한국음 이표기’는 수록자 중에서 오직 480자(표제어의 23.98%)에만 수록되어 있으며, 반절 위에 작게 표시된다. 「대한국문」에서 이 칸을 무엇을 의미하는지 따로 명시하지는 않으나 그 역할은 보기 (9)과 (10)으로 구분된다. 보기 (9)에 기재된 예시들에는 해당자의 이전 표기가 수록되어 있으며, 보기 (10)에 기재된 예시들에는 해당자에서 파생된 음가가 수록되어 있다. 두 용법 가운데 옛 표기를 나타내기 위해 쓰인 경우가 압도적으로 빈번하다. 특이한 점은 20세기 초에도 ・(아래아)가 빈번하게 쓰였음에도 (9')를 비롯한, ・가 중성에 오는 글자들은 전부 옛 표기로 간주했다는 것이다.

3. 영어 표기와 번역

영어 주해의 표기를 위한 지침으로서, <지석영본>은 「대한국문」 다음 장으로 「영국문(英國文)」을 기재한다. 「영국문」은 영문자의 정서(인쇄체), 초서(필기체), 해당하는 한글 음가를 소개하며, 그 옆에 영단어 발음 전사에 있어 지침을 세 줄로 요약한다. 이 세 지침은 보기 (11)과 같으며, 「대한국

4) 형태소 분리 표지 <ㅅ>이 과다분절 표기의 원인이라는 소견처럼 서술되어 있지만, 사실 빈번한 분절과 유음화를 금하는 과다한 두운 법칙은 <지석영본> 전체에 산발적으로 나타나는 특징이다. 예) 躡 - 쪽구릴 준 (四十三.1366) 得 - 엇'을 득 (四十八.1526) 驚 - 놀'날 경 (六十.1896)

문」의 ‘ㅅ’ 사용처럼 <지석영본>의 독자적인 표기법을 제시한다.

- (11) 一. 英音を 國文으로 難形한 字난 傍加圈標하니 ㄴ아씨의 類
 一. 英字問에 加書橫劃者난 上下句節의 接續과 連書하난 標
 一. 英字를 國文으로 繙譯한 中 特히 細小한 字난 核音を 有若無하게 做聲하난 標
- (12) 咽 - Throat - 으로트 (五.186) 醋 - Vinegar - 엑네까 (三十一.988)
 三 - Three - 쓰리 (六十一.1934) 女 - Female - ㅍ메일 (一.15)
- (13) 眉 - Eyebrow - 아이|부로우 (五.133) 醃 - Meatsauce - 미|트|쑈오쓰 (三十一.981)
 媼 - Old woman - 올드|우오만 (二.49) (cf. 翁 - Old man - 올드만 (二.48))
- (14) 鴨 - Duck - ㄷ크 (十六.491) 矇 - Ignorant - 익그노|란트 (三十七.1167)
 蜨 - Asp - 아습푸 (十九.599) 烟 - Smoke - 스목크 (十九.605)
- (15) 愛 - Love - 을노ㅅ (五十八.1848) (cf. 朽 - Rot - 으로트 (五十六.1760))

보기 (11)의 첫 번째 줄은 한글 체계로 전사하기 어려운 영어 음가를 위해 글자 초성 앞에 권표(圈標) ‘ㅇ’을 붙인 ㅇ계 합용병서를 사용함을 알려준다. <지석영본>에서 표기된 ㅇ계 합용병서는 총 다섯 종류로, ㄴ /θ, ð/, ㅅ /v/, ㅅ /z/, ㅅ /θ, ð/, ㅅ /f/로 나뉜다. 이 가운데 ㅅ은 순경음 ‘빙’음가와 동일하며, ㅅ은 반치음 ‘△’를 뜻한다. 기존 훈민정음 붕과 △ 대신 ㅇ계 합용병서 ㅅ과 ㅅ을 채택한 것은 특이사항인데, 본서에서 붕과 △은 화음(華音)의 /ㄷ/와 /ㄱ/를 표기하는 용도로만 쓰이기 때문이다. (‘4. 화음 전사 특징과 사성 표기 오류’ 참고.)

다섯 종류의 ㅇ계 합용병서 가운데 ㅅ는 「영국문」에서 Z 자의 영음(英音)을 소개할 때에만 보이며(“Z - 제트 ㅅ”) 본문에서는 쓰이지 않는다. 수록자의 유성 치경 마찰음 /z/는 ‘灰 - Ashes - 아췌쓰’(十.315)나 ‘紗 - Gauze - 짜우쓰’(二十八.888)같이 ㅅ계 병서으로 전사되기 때문이다. 나머지 ㅇ계 합용병서 ㄴ, ㅅ, ㅅ, ㅅ은 보기 (12)와 같이 영어 마찰음 전사에 쓰이며, <지석영본>에는 328개의 용례가 있다. 사용 빈도는 ㅅ (192개, 58.5%), ㅅ (76개, 23.2%), ㅅ (51개, 15.5%), ㅅ (9개, 2.7%) 순이었다. (이지현, 2018:866.) 다만 ㅇ계 합용병서와 달리 ㅅ계 합용병서는 된소리 전사 이상의 의미를 갖지 못한다 사료된다. 이는 15세기 말엽부터 ‘ㅅ’을 제외한 된소리는 각자병서가 아닌 합용병서로 표기되었고, 18세기 후반에 이르면 ‘ㅅ’, ‘ㅅ’, ‘ㅅ’, ‘ㅅ’으로 일반화되기 때문이다. ㅅ계 합용병서의 자음들은 본래 제각기 음가를 지닌 자음군(cluster)이었으나 ‘ㅅ’의 음가가 소실되고 된소리 표기 기능만 남았다. (우형식, 2014:149-152.)

보기 (11)의 두 번째 줄은 영단어 사이에 첨가하는 가로획(橫劃) ‘|’을 소개하며, 이 가로획은 두 단어를 이어서 발음함을 나타내는 표지로 기능한다. 보기 (13)에 기재된 전사 용례들에서, ‘|’이 ㉠ 단어 내 음절 구분, ㉡ 합성어(compound)로 결합시킨 형태소 간의 음절 구분, ㉢ 띄어쓰기로 구분된 단어들의 음절 구분의 역할을 수행함을 알 수 있다. <지석영본>에서는 임의의 합성어 또는 명사구를 표제어에 대한 번역어로 제시하는 경우가 많으나, ‘Old man - 올드만’(二.48)에서 보이듯이 ‘|’이 모든 합성어나 명사구 표기에 사용되는 것은 아니다.

마지막으로, 보기 (11)의 세 번째 줄은 세소(細小)한 글자에 대한 설명으로, 작게 표시된 글자의 음가를 있지만 없듯이 발성하라고 지시한다. 세소자는 두 용도로 사용되는데, 하나는 종성(coda)에 한 음소밖에 배치할 수 없는 한글 체계를 보완하기 위한 용도이다. 보기 (14)의 예시들처럼, 세소자는 단어 끝 또는 음절 끝에 배치되어 양음절에 걸친 이중 자음(ambisyllabic geminate)을 형성하되 종성이 두 음절로 완전히 분리되지 않도록 기여하는 역할을 한다.

하지만 <지석영본>의 세소자는 보기 (14)의 ‘Smoke - 스목크’(十九.605)처럼 초성(onset)의 음소들을 한 음절로 발음되도록 돕는 역할을 수행하지 않는다. 단어 초 또는 음절 초에 배치되는 글자는 설측음 /l/이 초성에 올 때 쓰는 세소자 ‘을’뿐이며, 그 예시들은 보기 (15)에 수록되어 있다. ‘을’의 받침 ‘ㄴ’을 강제로 종성에 둔 것은 측음으로 발음되도록 하는 표기 전략으로 간

무성 마찰음		유성 마찰음		유음	
/f/	/θ/	/v/	/ð/	/r/	/l/
ㅍ, ㅍ	ㅌ, ㅌ, ㅌ	ㅍ, ㅍ	ㅌ, ㅌ	Ø, ㄹ, 어초에서 /으ㄹ-/	Ø, ㄹ, 어초에서 /을ㄹ-/

주된다. (이석재, 2020:233.) 아울러 /l/이 세소자 ‘을’로 표기하는 한편, 권설음 /r/은 세소자가 아닌 ‘으’로 표기하여 두 음소가 다르다는 점을 강조했다.

그러므로 「영국문」과 상기 예시들에서 나타난 <지석영본>의 영단어 발음 전사 특징들은 상기 표와 같이 정리할 수 있다. 양식은 이석재(2020)를 참고하되 필자의 연구를 토대로 수정을 거쳤다.

- | | |
|--------------------------------------|--------------------------------------|
| (16) 兒 - *Dhid - 차일드 (一.1) | 鎧 - *Armmour - 암머 (二十九.934) |
| (17) 車 - Wheel - *웨일 (二十四.740) | 蜃 - Earthworm - *이아쯔우웜 (二十.600) |
| (18) 鮫 - *Porpoise - 포퍼쓰 (十八.584) | 蟪 - Cicada - *씨카다 (十九.603) |
| (19) 芎 - ?Scentherb - 쉐트 헐뿌 (十一.338) | 蘭 - ?Felicityherb - 엘니시티 헐부 (十一.332) |

아울러, <지석영본>에서의 영어 전사·번역 오류는 보기 (16)에서 (19)의 네 부류로 요약할 수 있다. 보기 (16)은 활자 오류 또는 철자 오인으로 인한 단순 영문 오타로, 본서에서 총 44건(2.20%)의 영문 오타가 발견되었다. 보기 (17)은 <지석영본>에서 가장 흔하게 보이는 전사 오류 유형이다.⁵⁾ 모음의 강세 유무를 오인하여 발생한 오류가 가장 빈번했으며, 다음으로 흔한 오류는 ‘嶼 - Island - 이슬난드’(八.247)를 비롯한, 영어 목음을 고려하지 않은 전사였다.

보기 (18)과 (19)는 단어의 번역에 관한 부류들을 나타낸다. 보기 (18)은 의미론적인 오류로, 복어(鮫)가 쇄돌고래(porpoise)로 오역되고 메뚜기(蟪)가 매미(cicada)로 오역되는 등의 미묘한 혼동으로 인한 오류가 보인다. 또한 보기 (19)는 오역보다는 의역에 가까운 예시들이다. ‘芎, 蘭’자를 각각 ‘angelica, orchid’로 번역하는 대신에 임의로 명사구 ‘scent herb, felicity herb’를 조어했다. <지석영본>에 만연한 오역과 의역은 일본어 주해에서도 나타나는데, 이에 대해서는 ‘5. 일본어 훈음 표기와 번역’과 ‘6. 영어·일본어 번역에서의 주안점들’에서 다룬다.

4. 화음 전사 특징과 사성 표기 오류

<지석영본>의 화음 운모 전사는 이미 황옥립(2021)이 체계적으로 다룬 바가 있으므로 본고는 <지석영본>의 사성(四聲)표기 오류와 권설음·순치음 표기에 초점을 맞춘다. <지석영본>의 중국어 전사 지침은 「대한국문」 바로 다음에 「화음(華音)」이라는 이름으로 수록되어 있는데, 그 일부를 발췌하면 보기 (20)과 같다.

- (20) 促而中曰上平① 長而輕曰下平② 曲而緩曰上聲③ 直而急曰去聲④
 華音之 수者난 수우之間音이요 부者난 부우之間音이니 輕唇音 吹唇而呼 (...) 音之輕微者난 傍加圈標
- | | | |
|---------------------------|----------------------|----------------------|
| (21) 儒 - 수 - ① (三.68) | 日 - 시 - ④ (六.164) | 仁 - 린 - ④ (三十二.1004) |
| (22) 夫 - 부 - ① (一.10) | 肺 - 예 - ① (五.441) | 奮 - 뽀 - ④ (六十二.1970) |
| (23) 時 - 췌 - ① (三十三.1041) | 左 - 췌 - ① (三十四.1068) | 草 - 뽀 - ① (十一.324) |

보기 (20)의 요점은 화음 사성과 /ㄱ, ㄷ/ 성모의 한글 전사 방법이다. <지석영본>은 원문자 ①, ②, ③, ④를 각각 상평(1성), 하평(2성), 상성(3성), 거성(4성)을 표시하는 성조 부호로 사용한다.

5) 오늘날의 『한국어 어문 규범』과 같은 현대의 영어 전사 정서법 기준을 20세기 초의 문헌에 소급하여 적용할 수는 없으므로, <지석영본> 내의 영어 전사 오류 개수는 관점에 따라 다르게 측정될 것이다. 본고에서는 측정된 전사 오류의 개수를 생략하는 바이다.

또한 /ㄱ/와 /ㄷ/은 각각 훈민정음 ‘ㄷ’, ‘ㄴ’으로 전사하며, ‘수’가 ‘수’와 ‘우’ 사이의 소리를, ‘부’가 ‘부’와 ‘우’ 사이의 소리를 뜻한다고 기록되어있다. 보기 (21)의 예시들(ㄱㅅ, ㄱ, ㄱㅅ)을 비롯하여, <지식영본>에 수록된 22개(표제어의 1.10%)의 /ㄱ/ 성모자는 모두 권설음이 아닌 받치음 ‘ㄷ’으로 표기되어 있다. 『A Chinese-English Dictionary』(1892)에서 정립된 웨이드-자일스 로마자 표기체계(Wade-Giles)에서 /ㄱ/를 프랑스어의 j[ʒ]와 유사하다 간주하여 ‘j’로 분류했듯이, <지식영본>의 ‘ㄷ’ 사용은 당대의 보수적인 관화 발음에 영향받았을 가능성이 크다.

「화음」은 성모 /ㄷ/를 경순음 ‘ㄴ’으로 전사함을 원칙으로 두지만, ‘가볍고 경미한 음에는 권표(圈標) ‘ㅇ’을 더한다’는 설명을 부연한다. 보기 (22)의 예시들(ㄷㅅ, ㄷㅅ, ㄷㅅ)을 비롯한, <지식영본>에 수록된 /ㄷ/ 성모자는 경순음 ‘ㄴ’ 또는 ㅇ계 합용병서 ‘ㄴ’으로 전사되는데, 어느 경우에 ‘ㄴ’ 또는 ‘ㄴ’을 사용하는지는 임의로 판단한 것으로 추정된다. 아울러 화음 전사에는 ‘ㄴ’ 이외에도 총 세 종류의 ㅇ계 합용병서가 쓰이는데, 보기 (23)에 용례가 있는 ㄴ /ㄷ/, ㄴ /ㄷ/, ㄴ /ㄷ/가 그것들이다. 특히 (명목상) 영어 마찰음 /z/에 대한 전사인 ‘ㄴ’이 화음에서는 권설음 성모 /ㄷ/과 대응하는 점은 이례적이다.

- | | | |
|---------------------------|-----------------------|------------------------|
| (24) 渠 - 취 - *ㄷ (八.233) | 駒 - 쥐 - *ㄷ (十七.536) | 飪 - 신 - *ㄷ (四十六.1445) |
| (24') 衾 - 친 - *ㄷ (三十.944) | 晦 - 휘 - *ㄷ (三十四.1066) | 逢 - 봉 - *ㄷ (五十八.1826) |
| (25) 棼 - *쑤 - *ㄷ (七.192) | 穀 - *구 - *ㄷ (二十八.891) | 癩 - *췌 - *ㄷ (四十一.1304) |

「아학편」은 한자를 표제어로 삼는 관계로, <지식영본>의 주요한 중국어 전사·번역 오류는 보기 (24)와 (25)로 요약이 가능하다. 보기 (24)는 성조 표기 오류로, 조사된 오류 136건(표제어의 6.80%) 가운데 98건(72.06%)에 해당했다. 단순 오기로 일부 존재하나 거의 모든 성조 오류들은 동일한 양상을 띄는데, 바로 수록자가 아닌 성부자(聲符字) 또는 해성자(諧聲字)의 성조를 적어둔 것이다. 보기 (24)의 예시들에는 성부자(汭[ㄴㅅ], 句[ㄴㅅ], 壬[ㄴㅅ])의 사성이 표기되었으며, 보기 (24')의 예시들에는 해성자(岑[ㄴㅅ], 悔[ㄴㅅ], 峰[ㄴㅅ])의 사성이 표기되었다.

보기 (25)는 음과 성조 모두에서 오류가 있는 경우들을 담는다. 이와 같은 오류들은 수록자를 유사한 뜻이나 형태의 한자와 오인하여 발생한다. ‘쑤’는 수록자 棼과 동일하게 ‘빛자루’라는 뜻을 가진 掃[ㄴㅅ]에 대한 표기이며, ‘후’는 수록자 穀과자와 유사하게 생긴 穀[ㄴㅅ] 자의 음가이다. 또한 ‘췌’는 성부자 閒(閒의 이체)[ㄴㅅ]에 대한 음가가 기입된 사례이다.

5. 일본어 훈음 표기와 번역

<지식영본> 원본은 표제어에 일본어 훈음 주해를 표기했음에도, 원판에 「대한국문」이나 「영국문」 같은 형식의 서문을 수록하지 않았다. 하지만 이후 1912년 일본 산세이도(三省堂) 출판사에 의해 『일어유해(日語類解)』, 『한어초보(韓語初步)』과 엮여 출판된 판에는 「일본국문(日本國文)」이라는 이름의 새로운 서문이 추가되었다. 보기 (26)는 「일본국문(日本國文)」에서 발췌한 것이다.

- (26) カキクケコ 五字난 他字下에 處하면 ㄱ終成을 生하야 上字에 付하야 發音하고 本字난 不失原音하니
 タカダ가 セキシ키 ハク학구 ムケ목게 ソコ속고 之類라
 ガグゲゴ 五字난 他字下에 處하면 ㅇ終成을 生하야 上字에 付하야 發音하고 本字난 變하야
 아이우에오가 되니 タガ다가 ハギ항이 サグ상우 나게낭에 마그망오 之類라
- (27) 赤 - アカシ - 약가시 (三十五.1102) 柳 - ヤナギ - 야나기 (十三.399)
 (27') 奴 - ヤッコ - 얏고 (三.66) 純 - モツバラ - 몯바라 (五十二.1534)

보기 (26)은 <지식영본>에서의 카행과 가행의 표기법을 다룬다. 보기 (27)의 전사 예시들처럼, <지식영본>은 카행 앞에 오는 자에 중성 ‘ㄱ’을 더하고, 가행 앞에 오는 자에 중성 ‘ㅇ’을 더하는 양음절 이중 자음으로써 무기음과 비탁음을 구현한다. 「일본국문」은 축음(ッ)을 ‘ㅅ’ 중성으로 전사함을 원칙으로 두나, 일부 축음은 보기 (27')의 ‘몯바라’처럼 양음절 이중 자음이 적용되기도 한다.

- (28) 思 - オモフ - 오모우 (五十八.1838) 想 - ヲモフ - 오모우 (五十八.1839)
 爾 - ナンジ - 난지 (六十二.1966) 汝 - ナンチ - 난지 (六十二.1967)
 (29) 裾 - モスソ - *모스조 (二十九.914) 異 沉 - シツム - *시쓰무 (五十二.1637)
 隸 - *シモ - 시모 (三.73) 舉 - *アグル - ʔ양구루 (三十八.1205)

또한 <지석영본>의 일본어 표기 특징은 보기 (28)처럼 역사적 가나 표기법과 신생 표기법이 혼재되어 있다는 점인데, 특히 탁점(・)이 일정하게 표기되어 있지 않다. 그러므로 집계된 일본어 전사 오류 130건⁶⁾ 가운데 대부분은 탁점과 관련되어 있다. 보기 (29)의 ‘모스조’는 ヌ자가 탁음이라 간주한 오표기이며, ‘시쓰무’는 반대로 ヌ자가 청음이라 간주한 오표기이다. 그 밖의 오류 유형으로는 시모(<シモベ>)를 비롯한 글자 누락이나 양구루(<アグル; アゲル의 잘못된 표기>)와 같은 글자 오인이 주되었다.

- (30) 陵 - ʔオホオカ - 오오옥가 (七.207) 註 - ʔアキラカニス - 아기라카니스 (三十九.1243)
 (31) 珠 - ʔカヒノタマ - 가이노다마 (十.300) 種 - ʔタネマク - 다네막구 (四十六.1428)

영어 표기와 마찬가지로, <지석영본>의 일본어 주해에서는 의역이 많이 집계되었다. 의역은 두 유형으로 나뉘는데, 하나는 보기 (30)과 같은 본래의 훈과 완전히 다른 의역으로 347건(표제어의 17.32%)이 집계되었다. 이러한 유형은 オホオカ(<大丘>)와 アキラカニス(<明キラカニス>)처럼 본래의 훈 대신 뜻을 설명하는 별도의 설명을 삽입한다. 또 다른 유형은 본래의 훈에 새로운 음을 더한 경우로 127건(표제어의 6.34%)이 집계되었다. 보기 (31)의 カヒノタマ(<貝ノ珠>)와 タネマク(<種蒔ク>)는 본래의 훈에 이해를 돕기 위한 추가 설명을 더했음을 알 수 있다.

오음	한음	오·한음 동일	관용음	분류 불가 ⁷⁾
156자 (7.79%)	1,088자 (54.32%)	625자 (31.20%)	45자 (2.25%)	89자 (4.44%)

- (32) 異 溺 - *ジャク - 작구 (五.159) 薺 - *ケイ - 케이 (十五.450)
 (33) 苗 - ビャウ - *묘 (十五.468) 韃 - ヒツ - *힌쓰 (三十.935)

<지석영본>의 일본어 음은 한음에 편중된 경향을 보이는데, 이는 한 표제어에 하나의 음만 기재할 수 있는 주해의 구성 때문이다. 일본 훈에서의 오류는 일본어 표기 오류인 보기 (32)와 한글 전사 오류인 보기 (33)으로 양분된다.

6. 영어·일본어 번역에서의 주안점들

<지석영본>의 영어·일본어 주해의 특징은 (편집자 전용규의 임의에 의한) 의역이 만연하다는 점이지만, 의역을 적용된 표기가 본래의 뜻에서 동떨어진 경우가 빈번하다는 문제가 있다.

- (34) 旬 - ʔトヲカ - Ten-days 望 - ʔジュゴニチ - Ides
 晦 - ミソカ - ʔDark 朔 - ツイタチ - Calends (三十四.1064 - 1067)
 (35) 芝 - ʔヒジリクサ - ʔFelicityherb 蘭 - アララギ - ʔFragrantplant
 蕙 - カホリクサ - ʔFragrantplant 菖 - ʔシャウフ - Calamus (十一.332 - 335)
 (36) 模 - ノリ - Model 楷 - ノリ - Model
 型 - ʔイカタ - Model 範 - ʔイカタ - Model (六十三.1988 - 1991)

보기 (34)에서 (36)는 모두 표제어에 영어·일본어 의역 주해를 포함한다. 보기 (34)는 의역(十日,

6) 일본음 전사 오류 64건, 일본훈 전사 오류 66건이 발견되었다. 1945년 이전의 문헌에 탁점을 표기하지 않는 현상은 매우 빈번했으므로, 단순히 탁점을 미표기한 경우는 오류로 간주하지 않았다.

7) 일본에서 쓰이지 않아서 읽음이 없는 자나, 본래의 음과 완전히 틀리게 표기된 경우는 ‘분류 불가’로 두었다.

十五日)이 표제어 구성(열흘-보름-그믐-초하루)에 대해 상대적으로 잘 풀이된 예시이다. 단 ‘Dark’는 날짜와 다르게 그믐달의 상에서 비롯한 의역으로, 표제어 구성에서 어긋난 주제어로 번역을 했다는 한계가 있다. 반면 보기 (35)는 의역이 표제어(지초-난초-혜초-창포)와 호응하지 않는 예시이다. 芝자의 일본어 훈 ヒジリクサ는 芝자가 본래 ‘잔디(シバ)’로 해석되기 때문에 설정한 의역으로 추정되며, 蕙자의 영어 주해 Fragrantplant는 일본어 훈 カホリクサ에 대한 직역이다.

한편 <지식영본> 번역의 일부는 표제어들 간의 구분을 고려하지 않아서 문제가 된다. 보기 (36)의 표제어들은 유의어 모해(模楷)와 형범(型範)으로 구성되어 있다. 글자 간 의미가 유사하지만, 무분별하게 Model과 イカタ(<良イ型)로 의역되어 있어서 표제어 간의 구분에 혼선을 빚는다.

- | | |
|--|--|
| (37) 兩 - ² ハカリノナ - *Tail | 鎰 - ² ハカリノナ - Catty-and-half (六十二.1962, 1963) |
| (38) 堯 - ² テイキャウ - ² Thyyo | 舜 - ² テイシュン - ² Thyshoon |
| 禹 - ² カウシ - ² Hawsi | 湯 - ユ - Soup (六十三.1996 - 1999) |
| (39) 孔 - アナ - Orifice | 孟 - ヲサ - Chief |
| 顔 - カホ - Phiz | 曾 - カツテ - Past (六十三.2000 - 2003) |

보기 (37)에서 (39)는 특이한 번역이 적용된 예시들이다. 보기 (37)의 양(兩)과 일(鎰)의 무게 단위 표제어에 대해, 일본어 훈에는 오직 ‘저울질의 단위(量リノ名)’이라고 뭉뚱그려 적혀있는 한편, 영어 주해에는 Tail(Tael 또는 Tahil의 오폭기)과 Catty-and-half라는 단위로 치환되어 있다.⁸⁾

보기 (38)는 요순우탕(堯舜禹湯)이 수록자들로 설정되어 있으며, 堯, 舜, 禹자에 대한 일본어 훈은 기존 훈에 부연 설명을 붙인 ティキャウ(<帝堯>, ティシュン(<帝舜>, カウシ(<夏禹氏)로 번역되어 있다. 또한 영어 주해 Thyyo, Thyshoon, Hawsi는 이들에 대한 전사이다. 하지만 본문의 구성에서 탕왕(湯王)을 뜻하는 湯자에는 다른 세 자와 달리 구성에 따른 의역이 적용되어 있지 않으며, ユ와 Soup는 모두 湯자의 본뜻 ‘탕(끓인 것)’을 의미한다.

보기 (39)는 『아학편』 본래의 구성 의도가 번역에 반영되지 못한 예시이다. 공맹안증(孔孟顔曾)은 유학의 네 성현인 각각 공자, 맹자, 안희, 증삼을 뜻하지만, 수록자들의 본뜻인 구멍(アナ, Orifice), 맷이(ヲサ, Chief), 얼굴(カホ, Phiz), 이전(カツテ, Past)으로 번역되었다.

7. 마치며

본고에서는 『아학편』 <지식영본>의 구성과 외국어 주에 대한 전사-번역 양상을 개략적으로 요약했으며, 선행 연구들의 결과에 새로이 조사한 바를 부연했다. 그 결과로, 본고에는 ㉔ 한국어 훈음 표기에서의 ‘, ㅅ’의 사용, ㉕ 영어 전사에서의 함용병서, 가로획, 세소자의 사용, ㉖ 화음 전사에서의 원문자와 ‘△, ㅂ’의 사용, ㉗ 일본어 전사에서의 양음절 이중 자음 사용, ㉘ 영어-일본어 주해에서의 의역을 <지식영본>의 특징으로 제시하였다.

국어 자료로써의 『아학편』은 예로부터 다방면에서의 연구가 이뤄져 왔다. 하지만 다국어 자전으로 써의 <지식영본>은 아직 적절한 요약이 없어서 연구를 진행하는 동안 아쉬울 따름이었다. 외국어 교육으로 구국을 피하고자 했던 지식영 선생의 큰 뜻을 헤량하기 위해, 향후에는 <지식영본>에 대한 연구가 다양한 언어의 관점에서 구축되기를 소망하는 바이다.

8) 한자문화권에서 근(斤)은 약 600g, 량(兩)은 약 37.5g (『표준국어대사전』에 의하면 37.7g)의 무게를 뜻하는 단위였으며, 斤은 catty(<말레이어 kati)로, 兩은 tael(<말레이어 tahil)로 음차되었다. 일(鎰)은 20량 또는 24량을 일컫는데, ‘1량 = 37.5g’ 등식으로 계산하면 ‘24량 = 900g = 1.5근 (∵ 600g = 1근)’이다. 그러므로 표제어 兩과 鎰에 대한 번역어 *Tail과 Catty-and-half는 정당하다고 여길 수 있다.

“A catty - comprising 16 taels (since the 16th century; in the 14th century it was 20 taels) - is traditionally equivalent to one pound avoirdupois, formalised as 604.78982 grams in Hong Kong, with similar values for Singapore’s and Malaysia’s kati; but rounded to 600g in Taiwan, Japan, Korea and Thailand.” (Lim, 2018. 밑줄은 필자에 의함.)

참고문헌

- 成慶 (2006). 『兒學編』과 『日語類解』의 日本漢字音의 淸濁音에 관해서, 동북아 문화연구, 11, 337-353.
- 우형식 (2014). 국어 된소리 표기법의 변천 영상. 우리말연구, 39, 141-179.
- 이석재 (2020). “아학편(兒學編)” 영단어 발음의 한글 표기에 관한 소고. 인문언어, 22(2), 201-228.
- 이준환 (2014a). 『아학편(兒學編)』의 구성과 뜻풀이의 특징에 대하여. 반교어문학회, 36, 29-64.
- 이준환 (2014b). 池錫永 『兒學編』 영어 어휘의 한글 표기와 국어의 음운론적 대응 양상 -자음을 대상으로 하여-. 국어사연구, 18, 249-292.
- 이지현 (2018). 지석영 편(1908), 《아학편(兒學編)》에 나타난 훈민정음 합용병서를 활용한 발음교육 효과연구. 문화와융합, 40(8), 855-876.
- 林煥澤 (1998). 特輯 - 湖南의 학문전통과 漢文學: 丁若鏞의 康津 流配時의 敎育활동과 그 성과. 한국한문학연구, 21(0), 113-150.
- 정재영 (2000). 茶山の 《兒學編》, 국어사 자료 연구 창간호, 국어사학회, 15, 1-174.
- 황영순 (2014). 한국에서의 미국영어 교육의 변천과정 조망. 미국사연구, 40, 201-238.
- 황옥림 (2021). 지석영 『兒學編』(1908) 한글 주음 ‘華音’의 韻母에 대하여. 언어, 46(3), 859-895.
- Lim, Lisa (2018). How catty and tael entered the English language, along with picul, mace and candareen - Asian weight measurements. South China Morning Post. 18 Jan, 2018.

On the Diverse Paradigm of Rising Declaratives in English

Junseon Hong
Seoul National University
jshong512@snu.ac.kr

KACL-KASELL Spring Conference
June 3, 2023

1

Introduction

- Assertions and questions are basic direct speech acts.
 - In English, they canonically align with distinct syntactic forms.
 - Assertions correspond to declaratives as in (1a).
 - Questions correspond to polar interrogatives as in (1b).
- (1) a. She's home[↓]. (Falling) Declarative
b. Is she home[↑]? (Rising) Polar Interrogative
- However, (2), which involves a declarative sentence with rising intonation, conveys a marked meaning.
- (2) She's home[↑]? Rising Declarative (RDs)

2

Introduction

Purpose of the Study

- To shed some light on specific functions of RDs that have not received as much attention
- To propose how the interpretive effects of RDs arise from the combination of their semantics and pragmatics, which allows them to be interpreted as one function over the others

3

Paradigm

Rising Declaratives

- RDs that overlap with assertions are referred as Assertive Rising Declaratives (ARDs; Jeong, 2018).

(3) Assertive Rising Declaratives

A: Where's Sally?

B: (Um...) *She's home?*

- RDs that overlap with questions are referred as Inquisitive Rising Declaratives (IRDs; Jeong, 2018).

(4) Inquisitive Rising Declaratives

A: (flight agent) There's one flight to Seoul.

B: (customer) *The flight leaves at 10am?*

4

Paradigm

Assertive Rising Declaratives: Epistemic Uncertainty and Metalinguistic Uncertainty

- ARDs can be further categorized into two, in terms of the uncertainty they implicate.
- Epistemic Uncertainty ARDs implicate the tentativeness on to the truth value of the expressed proposition.

- (5) A: Where's Sally?
 B: (Um...) *She's home?*
 B': (Um...) She **might** be home.

- Metalinguistic Uncertainty ARDs implicate the tentativeness on the relevance to the current Question Under Discussion (QUD; Roberts, 1996, 2012).

- (6) A: Do you speak Chinese?
 B: *I speak Cantonese?*
 B': #I **might** speak Cantonese.

5

Paradigm

Assertive Rising Declaratives: Politeness

- Both subtypes of ARDs can be used as a politeness strategy as they are more indirect than canonical assertions (Jeong, 2021).

- (7) A: Do you want a glass of water?
 B: *I'll have a beer?*

- (8) [Context: A waiter **A** is introducing himself.]
 A: *Hello, my name is David? I'm your water today?*

6

Paradigm

Inquisitive Rising Declaratives: Confirmative and Contradictory

- IRDs encompass confirmative and contradictory functions.
- Confirmative IRDs implicate the speaker's bias toward the expressed proposition.

(9) A: (flight agent) There's one flight to Seoul.
B: (customer) *The flight leaves at 10am?*

- Contradictory IRDs implicate the speaker's bias toward the negation of the expressed proposition.

(10) A: I went to the concert last night. Dave is a good singer.
B: *Dave is a good singer?* No way. You must be thinking about John.

7

Paradigm

Inquisitive Rising Declaratives: Mirative

- RDs can express mirativity (Goodhue et al., 2015).
- Mirative IRDs are a specific subtype of Contradictory IRDs.
 - To express her surprise, the speaker should not have been committed to p prior to the discourse act.

(11) A: I met Dave's brother yesterday.
B: (What?) *Dave has a brother?*

8

Preliminaries

- I adopt the expanded Table model (Farkas & Bruce, 2010; Malamud & Stephenson, 2015).

(12) Discourse Components

- a. COMMON GROUND (CG):
the set of propositions that all speakers are publicly committed to
(Stalnaker, 1978)
- b. DISCOURSE COMMITMENT (DC_x):
the set of propositions that the speaker has publicly committed to during the conversation up to the relevant time, and which are not shared by all the other participants
(Farkas & Bruce, 2010)
- c. TABLE (T):
the stack that records at-issue content in the conversation
(Farkas & Bruce, 2010)

9

Preliminaries

(12) Discourse Components

- d. PROJECTED COMMON GROUND (CG^*):
the set of potential CG s that gives possible resolutions for the top issue on the Table in the next expected stage of the conversation
(Farkas & Bruce, 2010; Malamud & Stephenson, 2015)
- e. PROJECTED DISCOURSE COMMITMENT (DC_x^*):
the set of propositions that the speaker is expected to become committed to (Speaker's Projected Discourse Commitments; DC_{sp}^*) or the best guess of commitments made by other interlocutors (Addressee's Projected Discourse Commitments; DC_{ad}^*)
(Malamud & Stephenson, 2015)
- f. PROJECTED TABLE (T^*):
the stack that records at-issue content which the speaker expects to be relevant to the current Question Under Discussion (QUD)
(c.f., Malamud & Stephenson, 2012; Bhadra, 2020)

10

Contributions of Rising Intonation

(13) Conventional Discourse Effect (CDE) of Declaratives (Farkas & Bruce, 2010; Farkas, 2022)

$$(i) Table_o = Table_i \cup \{p\}$$

$$(ii) DC_{sp,o} = DC_{sp,i} \cup p$$

- I propose that rising intonation overrides the convention of falling declaratives in two distinct ways.
 1. increasing the inquisitive content of the proposition → Semantic Convention
 2. projecting discourse components → Pragmatic Reasoning

11

Interpretation Semantic Convention

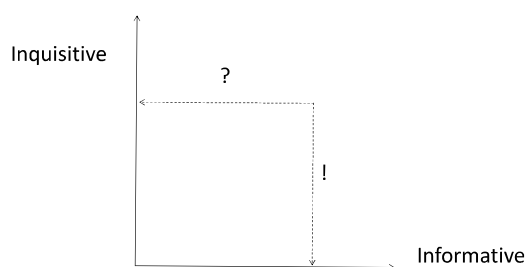


Figure 1 Informative Content and Inquisitive Content

- In the framework of Inquisitive Semantics (Ciardelli et al., 2019), a sentence not only conveys informative content but it also expresses inquisitive content by raising an issue.
 - Falling declaratives are non-inquisitive by default.
 - Rising polar interrogatives are basically non-informative with informative content being trivialized.

12

Interpretation

Semantic Convention

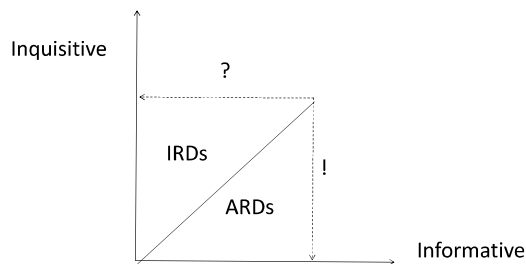


Figure 2 Informative Content and Inquisitive Content of Rising Declaratives

- ARDs: informative > inquisitive (a singleton set $\{p\}$) \approx Falling Declaratives (Hamblin, 1971)
- IRDs: Informative < Inquisitive (a non-singleton set $\{p, \neg p\}$) \approx Polar Interrogatives (Karttunen, 1977).

13

Interpretation

Semantic Convention

- The convention of two RDs derives from their steepness of their intonation.

Assertive Rising Declaratives	Inquisitive Rising Declaratives
Weak Rise	Steep Rise

Jeong (2018)

- A weak rise ($H^* H-H\%$) increases inquisitive content up to the point where it is no higher than informative content.
 - The proposition remains primarily informative, resulting in a singleton set $\{p\}$.
- A steep rise ($L^* H-H\%$) increases inquisitive content to surpass informative content.
 - The content is shifted from a singleton set $\{p\}$ to a non-singleton set $\{p, \neg p\}$, turning the primary speech act into a question.

14

Interpretation: ARDs

Pragmatic Reasoning

- An ARD that conveys a proposition that is a subset of the current QUD is construed as an Epistemic Uncertainty ARD.

(14) A: Where's Sally?
B: (Um...) *She's home?*

- The content of an ARD, *{Sally is home}*, is a subset of the current QUD, *{Sally is home, Sally is at school, Sally is at the café, Sally is at the gym, ...}* (i.e., $p \in P$).

15

Interpretation: ARDs

Pragmatic Reasoning

- An ARD that conveys a proposition that is NOT a subset of the current QUD is construed as a Metalinguistic Uncertainty ARD.

(15) A: Do you speak Chinese?
B: *I speak Cantonese?*

- The content of an ARD, *{I speak Cantonese}*, is not a subset of the current QUD, *{I speak Chinese, I don't speak Chinese}* (i.e., $p \notin P$).

16

Discourse Effects: ARDs

(16) Discourse Effect of Epistemic Uncertainty ARDs (updating c_i with $\{p\}$)

(i) $Table_o = Table_i \cup \{p\}$

(ii) $DC_{sp,o}^* = DC_{sp,i}^* \cup p$

→ *Commitment* is projected to the DC_{sp}^* : Speaker's uncertainty on the truth value

(17) Discourse Effect of Metalinguistic Uncertainty ARDs (updating c_i with $\{p\}$)

(i) $Table_o^* = Table_i^* \cup \{p\}$

(ii) $DC_{sp,o} = DC_{sp,i} \cup p$

→ *Table* is projected to the $Table^*$: Speaker's uncertainty on the relevance to the QUD

(18) CDE of Declaratives (updating c_i with $\{p\}$)

(i) $Table_o = Table_i \cup \{p\}$

(ii) $DC_{sp,o} = DC_{sp,i} \cup p$

17

Discourse Effects: ARDs

- The politeness is analyzed in terms of the speaker's updating p to the projected components for each subtype of ARDs which result in enhancing the addressee's face.
 - Epistemic Uncertainty ARDs: the DC_{sp}^*
 - Metalinguistic Uncertainty ARDs: the $Table^*$

18

Interpretation: IRDs

Pragmatic Reasoning

- An IRD that is used in the context WITHOUT contextual evidence on addressee's belief or bias is attested as a Confirmative IRD.

(19) A: (flight agent) There's one flight to Seoul.

B: (customer) *The flight leaves at 10am?*

- No evidence in the context directly supports that the addressee would believe the flight leaves at 10am.

19

Interpretation: IRDs

Pragmatic Reasoning

- An IRD that is used in the context WITH contextual evidence on addressee's belief or bias is attested as a Contradictory IRD.

(20) A: I went to the concert last night. Dave is a good singer.

B: *Dave is a good singer?* No way. You must be thinking about John.

- The addressee's belief on *p* is known to the speaker from the explicit expression.

20

Discourse Effects: IRDs

(21) Discourse Effect of Confirmative IRDs (updating c_i with $\{p, \neg p\}$)

(i) $Table_o = Table_i \cup \{p, \neg p\}$

(ii) $DC_{sp,o}^* = DC_{sp,i}^* \cup p$

→ *Commitment* is projected to the DC_{sp}^* : Speaker's positive bias

(22) Discourse Effect of Contradictory IRDs (updating c_i with $\{p, \neg p\}$)

(i) $Table_o = Table_i \cup \{p, \neg p\}$

(ii) $DC_{ad,o}^* = DC_{ad,i}^* \cup p$

→ *Commitment* is projected to the DC_{ad}^* : Speaker's negative bias

(23) CDE of Declaratives (updating c_i with $\{p\}$)

(i) $Table_o = Table_i \cup \{p\}$

(ii) $DC_{sp,o} = DC_{sp,i} \cup p$

21

Discourse Effects: IRDs

- Mirative IRDs are modeled as illocutionary not-at-issue content (Rett, 2021) which updates flavored commitment to the DC_{sp} (Rett & Sturman, 2021).

(24) Discourse Commitments

Let DCa be sets of propositions of the form **is-surprised $_a(p)$** , representing the public commitments of a with respect to a discourse in which a and b are the participants, where:

- a. is-surprised $_a(p)$** is a public commitment of a iff ' a is surprised with p ' is a mutual belief of a and b .

(Rett, 2021, p. 326)

(25) Discourse Effect of Mirative IRDs (update c_i with $\{p, \neg p\}$)

(i) $Table_o = Table_i \cup \{p, \neg p\}$

(ii) $DC_{ad,o}^* = DC_{ad,i}^* \cup p$

(iii) $DC_{sp,o} = DC_{sp,i} \cup \text{is-surprised}_{sp}(p)$

22

Conclusion

		ARDs		IRDs	
		EU	MU	Conf.	Cont.
Semantics	Rising Intonation	Weak (H* H-H%)		Steep (L* H-H%)	
Pragmatics	Subset of the QUD	YES	No	-	
	Contextual Evidence	-		No	YES
Semantic Content		{p}		{p, -p}	
Projection		DC_{sp}^*	<i>Table</i> *	DC_{sp}^*	DC_{ad}^*

23

Selected References

- Bhadra, D. (2020) The semantics of evidentials in questions. *Journal of Semantics*, 37(3), 367-423.
- Ciardelli, I., Groenendijk, J., & Roelofsen, F. (2019). *Inquisitive semantics*. Oxford: Oxford University Press.
- Farkas, D. F. (2022). Non-intrusive questions as a special type of non-canonical questions. *Journal of Semantics*, 39(2), 295-337.
- Farkas, D. F., & Bruce, K. B. (2010). On reacting to assertions and polar questions. *Journal of Semantics*, 27(1), 81-118.
- Goodhue, D., Harrison, L., Su, Y. T. C., & Wagner, M. (2015). Toward a bestiary of English intonational tunes. In C. Hammerly & B. Pickett (Eds.), *Proceedings of the 46th Conference of the North East Linguistic Society (NELS)* (pp. 311-320).
- Hamblin, C. L. (1971). Mathematical models of dialogue. *Theoria*, 37(2), 130-155.
- Jeong, S. (2018). Intonation and sentence type conventions: Two types of rising declaratives. *Journal of Semantics*, 35(2), 305-356.
- Jeong, S. (2021). Deriving politeness from an extended Lewisian model: The case of rising declaratives. *Journal of Pragmatics*, 177, 183-207.
- Karttunen, L. (1977). Syntax and semantics of questions. *Linguistics and Philosophy*, 1(1), 3-44.
- Malamud, S. A., & Stephenson, T. (2012). Three ways to avoid commitments: The pragmatics of declarative force modifiers. In A. A. Guevara, A. Chernilovskaya, & R. Nouwen (Eds.), *Proceedings of Sinn und Bedeutung* 16(2) (pp. 403-416).
- Malamud, S. A., & Stephenson, T. (2015). Three ways to avoid commitments: Declarative force modifiers in the conversational scoreboard. *Journal of Semantics*, 32(2), 275-311.
- Rett, J. (2021). The semantics of emotive markers and other illocutionary content. *Journal of Semantics*, 38(2), 305-340.
- Rett, J., & Sturman, B. (2021). Prosodically marked mirativity. In D. K. E. Reisinger & M. Huijsmans (Eds.), *Proceedings of 37th West Coast Conference in Formal Linguistics (WCCFL)* (pp. 1-20). Somerville, MA: Cascadia Proceedings Project.
- Roberts, C. (1996). Information structure in discourse: Towards an integrated formal theory of pragmatics. In J. H. Yoon & A. Kathol (Eds.), *Papers in Semantics (Working Papers in Linguistics 49)*. Columbus, OH: The Ohio State University.
- Roberts, C. (2012). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5: 6, 1-69.
- Stalnaker, R. (1978). Assertion. In P. Cole (Ed.), *Syntax and semantics* (pp. 315-332). New York, NY: Academic Press.

24

디지털 네이티브 세대 그리고 초개인화 영어교육 생성형 AI 기반의 난이도 구분 지문 및 문항 생성 렉스퍼

Ver. 20230603A

Copyright 2018-2023 LXPER Inc. All Rights Reserved

2

FOUNDER & CEO



영어교육 및 자연어 처리 전문가
이 형 중 (Jason Lee)

20년 이상의 영어교육 경력과 5년간의 AI 연구를 바탕으로
디지털 네이티브 세대의 초개인화 영어교육을 실현합니다.

- 연세대학교 영어교육 석사 (학위논문: 한국 대학 신입생 영작문에 나타난 오류 연구)
- 코퍼스 언어학 기반 DB 구축 및 AI 기반 영단어 분석 서비스 LXPER AI엔진 개발 총괄
- AI 기반 개인 맞춤 영어 시험 문제 제작 서비스 AI TEST MAKER 기획, AI엔진 개발 및 콘텐츠 총괄
- 자연어 처리 관련 해외 논문 총 4편 (EMNLP2021 메인 컨퍼런스 포함) 공저 및 12건 이상의 AI 기술 특허 출원 · 발명인
- 2020 콘텐츠 스타트업 엑셀러레이팅 데모데이 최우수상 (대구콘텐츠기업육성센터 주관)
- 총 195권 교재 집필 경력, 23권의 고등 영어 교재 대표 저자(베스트셀러 10권), 172권의 초/중등 영어 교재 원고 총괄*

고등/수능 교재



2019년 베스트셀러 선정
고등수험서부문 (YES24 기준 '19.06)



2020년 베스트셀러 선정
EBS 변형문제집 영어부문 (네이버 기준 '20.05)

초등 교재



중등 교재



회사 정보

회사명 / 대표 주식회사 렉스퍼 LXPÉR Inc. / 이 형 중 Jason Lee (Founder & CEO)

주요연혁 **2018.12** 법인 설립

2018 - 2020 KOTRA지정 신규 수출화 기업 선정

2019.12 기술보증기금 제5회 벤처캠프 최우수 기업 선정(10억 보증)

2019.12 벤처기업 인증

2020.02 콘텐츠 스타트업 엑셀러레이팅 데모데이 1위 (최우수상, 대구콘텐츠기업육성센터 주관)

2020.04 기업부설연구소 인정

2020.07 중소벤처기업부 4차산업혁명 인공지능부문 정부과제 선정

2020.08 비상교육 + LXPÉR엔진 문제생성 교재개발 계약

2020.08 LXPÉR INDEX 해외논문출판 International Journal of Advanced Computer Science and Applications, 11(8), 2020

2020.10 LGCNS AI Tutor + LXPÉR엔진 API서비스 연동계약

2020.12 LXPÉR INDEX 2.0 해외논문출판 Association for Computational Linguistics

2021.04 AI기반 개인 맞춤형 영어 문제 자동 생성 플랫폼 AI TEST MAKER 출시

2021.11 해외논문출판 EMNLP2021 (main conference) Empirical Methods in Natural Language Processing, 11, 2021

2022.02 LXPÉR X 비상교육, 초중등 전국영어능력평가 업무 제휴 계약(총 5 레벨)

2022.04 XX교육, 초중등 듣기 학습 및 교수 콘텐츠 제작 및 공급 계약

2022.07 LXPÉR X 비상교육, 초중등 전국영어능력평가 VPEAT 1차시험 진행

2023.05 현재 AI기반 1:1 개인 맞춤 문법 학습 솔루션 Grammar PEEP 출시

2023.06 현재 LXPÉR X 비상교육, 초중등 전국영어능력평가 VPEAT 기본 수험서 제작 중

2023.06 현재 LXPÉR X ADB(아시아 개발 은행)와 방글라데시 공교육 진출 관련 세부 논의 중

*2023. 06. 03. 현재 기준

Personalized, AI-based English Education Platform Company

LEXICON EXPERT. LXPÉR.com

지문 입력

As you progress in your English studies, you'll find that personalized language instruction becomes increasingly significant. This approach recognizes that each student's language learning journey is unique, shaped by individual strengths, areas for improvement, and personal interests. Personalized instruction allows for learning at your own pace, focusing on areas that may be particularly challenging or intriguing to you. Whether it's mastering complex grammar rules or expanding your vocabulary to express yourself more precisely, a tailored learning approach helps you to achieve these goals more effectively. It ensures that your learning experience is not only educational but also engaging and personally rewarding. Each English phrase you master, every grammatical concept you grasp, are stepping stones on your path to becoming a proficient English speaker and writer. With personalized instruction, you're not just learning English—you're sculpting your own linguistic journey.

990/1000자

초기화

어휘 분석

어휘 분석 결과

As you progress in your English studies, you'll find that personalized language instruction becomes increasingly significant. This approach recognizes that each student's language learning journey is unique, shaped by individual strengths, areas for improvement, and personal interests. Personalized instruction allows for learning at your own pace, focusing on areas that may be particularly challenging or intriguing to you. Whether it's mastering complex grammar rules or expanding your vocabulary to express yourself more precisely, a tailored learning approach helps you to achieve these goals more effectively. It ensures that your learning experience is not only educational

난이도

☒ 초등기초 포함

문장순

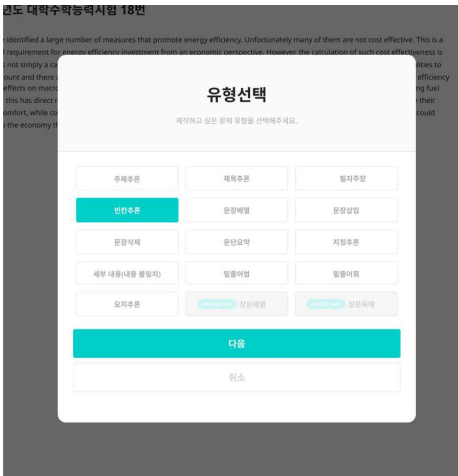
초등기초		초5~중2		중2~중3		중3~고2		고2~고3		대학(원)	
Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent
38	40%	15	16%	21	22%	16	17%	3	3%	1	1%

어휘 목록

등급	단어	뜻
★	as	~할 때, ~함에 따라, ~때문에, ~처럼, ~로(서), ~만큼
★	you	너, 당신, 너희, 여러분
★★	progress	진보하다, 발전하다
★	in	~에, ~안에
★	your	너의, 당신의, 여러분의
★	English	잉글랜드의, 영어의
★	study	공부, 연구
	you'll	you will의 줄임말
★	find	찾다, 발견하다, 알아보다
★	that	저것, 저
★	personalized	개인맞춤의, 개별화된, 이름이 들어간
★	language	말, 언어
★★	instruction	설명, 지시, 교본

생성형 AI의 시작. AI Test Maker(ATM) atm.lxper.ai

2021년 4월. AI기반 수능 및 내신 대비 문제 제작 서비스 출시



- **문제 유형을 선택한 후** 시험의 특성에 맞는 문제(내신형, 수능형)를 손쉽게 제작
- **LXPER AI가 매력적인 오답 자동 생성** 이후 학습자 수준 고려하여 직접 선별 및 수정 가능
- 기존 제작 방식 대비 문제 제작에 투입되는 시간 **최대 81.5% 감소**
- 문제 제작 경험이 많지 않은 비전문가도 인공지능의 도움으로 **전문성 있는 문제 제작** 가능

Personalized, AI-based English Education Platform Company

그리고 역사적인 2022년 11월 30일(한국 12월 1일)



생성형 AI를 공교육에 바로 활용가능한가?

Model: GPT-4

LK You are writing an English textbook for elementary school students in South Korea. Write a paragraph on the importance of personalized language instruction. It is most important that the vocabulary and grammar level be understandable to Korean elementary school students.

AI Learning English is a wonderful adventure and every adventure is unique, just like you! Personalized language instruction is very important because we all learn in different ways. It means having lessons that are made just for you! This way, you can learn new words and

초등기초	초5-중2	중2-중3	중3-고2	고2-고3	대학(원)
Count	Percent	Count	Percent	Count	Percent
81	79%	12	11%	7	6%
0	0%	2	1%	0	0%

LK You are writing an English textbook for middle school students in South Korea. Write a paragraph on the importance of personalized language instruction. It is most important that the vocabulary and grammar level be understandable to Korean middle school students.

AI Studying English becomes more exciting when lessons are tailored to your unique needs and interests, and this is what we call personalized language instruction. This approach is extremely valuable because it considers your individual learning style, pace, and goals. With lessons designed especially for you, you can focus on mastering the areas that challenge you the most. Remember, each new word you learn, each grammar rule you understand, they are the bricks that build your English language skills. With personalized instruction, you're not only studying English, you're shaping your own language adventure!

초등기초	초5-중2	중2-중3	중3-고2	고2-고3	대학(원)
Count	Percent	Count	Percent	Count	Percent
78	67%	16	13%	3	2%
0	0%	0	0%	0	0%

LK You are writing an English textbook for high school students in South Korea. Write a paragraph on the importance of personalized language instruction. It is most important that the vocabulary and grammar level be understandable to Korean high school students.

AI As you progress in your English studies, you'll find that personalized language instruction becomes increasingly significant. This approach recognizes that each student's language learning journey is unique, shaped by individual strengths, areas for improvement, and personal interests. Personalized instruction allows for learning at your own pace, focusing on areas that may be particularly challenging or intriguing to you. Whether it's mastering complex grammar rules or expanding your vocabulary to express yourself more precisely, a tailored learning approach helps you to achieve these goals more effectively. It ensures that your learning experience is not only educational but also engaging and personally rewarding. Each English phrase you master, every grammatical concept you grasp, are stepping stones on your path to becoming a proficient English speaker and writer. With personalized instruction, you're not just learning English—you're sculpting your own linguistic journey.

초등기초	초5-중2	중2-중3	중3-고2	고2-고3	대학(원)
Count	Percent	Count	Percent	Count	Percent
71	51%	19	13%	10	7%
0	0%	0	0%	0	0%

Personalized, AI-based English Education Platform Company

Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent
81	79%	12	11%	7	6%	0	0%	2	1%	0	0%

Learning English is a wonderful adventure and every adventure is unique, just like you! Personalized language instruction is very important because we all learn in different ways. It means having lessons that are made just for you! This way, you can learn new words and grammar at your own pace, and focus on what you are most interested in. This makes learning English more fun and easier. Just remember, every time you learn something new in English, you are adding a valuable piece to your language puzzle. And when all the pieces fit together, you'll be able to speak and write English confidently!

생성형 AI를 공교육에 바로 활용가능한가?

[illegible]

LXPER INDEX ???



초격차 기술 확보를 위해 항상 연구하는 LXPER

2020년 8월 LXPER Index를 시작으로 2020년 12월 LXPER Index 2.0을 거쳐
2021년 11월 자연어 처리 세계 최고 권위 학회인 EMNLP에서 메인 컨퍼런스 발표하였고
2023년 5월 EACL 2023 논문 출판까지 AI 기술 연구는 끊임없이 진행 중입니다.

LXPER Index: a curriculum-specific text readability assessment model for EFL students in Korea

Bruce W. Lee^{1,2}
¹Research & Development Center
LXPER, Inc.
Seoul, South Korea

Jason Hyung-Jong Lee¹
²Department of Computer & Information Science
University of Pennsylvania
Philadelphia, PA, USA

Abstract—Automatic readability assessment is one of the most important applications of Natural Language Processing (NLP) in education. Since automatic readability assessment allows the fast selection of appropriate reading material for readers at all levels of proficiency, it can be particularly useful for the English education of English as Foreign Language (EFL) students around the world. However, most readability assessment models are developed for the native readers of English and have low accuracy for texts in non-native English Language Training (ELT) curriculum. We introduce LXPER Index, which is a readability assessment model for non-native EFL readers in the ELT curriculum of Korea. We also introduce the Text Corpus of the Korean ELT Curriculum (CaKEC-text), which is the first collection of English texts from a non-native country's ELT curriculum with each text's target grade level labeled. In addition, we assembled the Word Corpus of the Korean ELT Curriculum (CaKEC-word), which is a collection of words from the Korean ELT curriculum with word difficulty labels. Our experiments show that our new model, trained with CaKEC-text, significantly improves the accuracy of automatic readability assessment for texts in the Korean ELT curriculum. The methodology used in this research can be applied to other ELT curricula around the world.

Keywords—Natural Language Processing; Machine Learning; Text Readability Assessment; EFL education

I. INTRODUCTION
Readability Assessment helps quantify the level of difficulty that a reader experiences in comprehending a certain text. Since automatic readability assessment enables the convenient selection of appropriate reading material for readers with different levels of proficiency [1], readability assessment has been an important field of research since as early as the 1980's [11]. Since then, more than 200 readability formulas were developed [3], but most of them concentrated on the general audience in the United States. We argue that there is a need for the development of an improved text readability assessment method for use in English as Foreign Language

LXPER Index 2.0: Improving Text Readability Assessment for L2 English Learners in South Korea

Bruce W. Lee^{1,2}
¹Dep. of Computer & Information Science¹
University of Pennsylvania
PA, USA
brucelee@seas.upenn.edu

Jason Hyung-Jong Lee²
²Research & Development Center²
LXPER, Inc.
Seoul, South Korea
jasonlee@lpxper.com

Abstract
Developing a text readability assessment model "specifically" for L2 English learners in South Korea has never had much attention in the field of Natural Language Processing. Hence, most developed models show extremely low accuracy for L2 English texts, up to the point where not many even serve as a fair comparison. In this paper, we investigate a text readability assessment model for L2 English learners in Korea. In accordance, we improve and expand the Text Corpus of the Korean ELT Curriculum (CaKEC-text). Each text is labeled with its target grade level. We train our model with CaKEC-text and significantly improve the accuracy of readability assessment for texts in the Korean ELT curriculum.

1 Introduction

Text readability assessment has been an important field of research since the 1940s. However, most research focused on the native audience in English speaking countries (Benjamin, 2012). In China, Japan, and Korea, many high and middle school students attend English language schools, in addition to their regular school classes. English subject plays an important role in the educational systems of these countries (McKay, 2003).

Despite the importance put in English education, the previous text readability assessment models have not been in active use in the three countries. This is due to the poor performance of traditional readability assessment models on L2 texts. We believe there is an immediate need for the development of an improved text readability

of linguistic features, consisting of the generic features of a text like total words, total sentences, and total syllables (Kincaid et al., 1975). Such features are effective predictors of a text's readability, but more curriculum-specific features are required for L2 text readability assessments. The key distinction between native readability assessment and L2 readability assessment is that L2 students rigorously follow the specific national ELT curriculum. Unlike native students who learn English from a variety of sources, most L2 students have limited exposure to English. In this research, we reduce the average assessment error by implementing some curriculum-specific features.

The contributions of this paper are: (1) we utilize and expand CaKEC-text, one of the few graded corpora with texts from an actual L2 curriculum; (2) we investigate novel linguistic features that were rarely used on an L2 corpus; (3) we evaluate our model against other readability models, show significantly improved accuracy, and prove that "grades" are better modeled using logistic regression, not linear regression.

2 Related Work

Research efforts in developing automated text readability assessment models for L2 students only emerged in the 2000s (Xu et al., 2016). Halloran et al. (2007) showed that grammatical features and lexical features play particularly important roles in L2 text readability prediction. Meanwhile, Vigliani and Meurers (2018) showed

Pushing on Text Readability Assessment: A Transformer Meets Handcrafted Linguistic Features

Bruce W. Lee^{1,3}
¹Univ. of Pennsylvania¹
PA, USA
brucelee@seas.upenn.edu

Yoo Sang Jang^{2,3}
²Univ. of Wisconsin-Madison²
WI, USA
yjang43@wisc.edu

Jason Hyung-Jong Lee³
³LXPER AI3
Seoul, South Korea
jasonlee@lpxper.com

Abstract

We report two essential improvements in readability assessment: 1. three novel features in advanced semantics and 2. the timely evidence that traditional ML models (e.g. Random Forest, using handcrafted features) can combine with transformers (e.g. RoBERTa) to augment model performance. First, we explore suitable transformers and traditional ML models. Then, we extract 255 handcrafted linguistic features using self-developed extraction software. Finally, we assemble those to create several hybrid models, achieving state-of-the-art (SOTA) accuracy on popular datasets in readability assessment. The use of handcrafted features help model performance on smaller datasets. Notably, our RoBERTa-REF-T hybrid achieves the *near-perfect classification accuracy of 99%*, a 20.3% increase from the previous SOTA.

1 Introduction

The long quest for advancing readability assessment (RA) mostly centered on handcrafting the linguistic features that affect readability (Pike and Newkirk, 2008). RA is a time-honored branch of natural language processing (NLP) that quantifies the difficulty with which a reader understands a text (Feng et al., 2019). Being one of the oldest systematic approaches to linguistics (Collins-Thompson, 2014), RA developed various linguistic features. These range from simple measures like the average count of syllables to those as sophisticated as semantic complexity (Buchanan et al., 2001).

Perhaps due to the abundance of disposable lin-

with BERT (Devlin et al., 2019), ELMo (Peters et al., 2018), Word2Vec (Mikolov et al., 2013), and GloVe (Pennington et al., 2014) embeddings showed only ~1% accuracy improvement from a single SVM model developed by Xia et al. (2016).

Even though deep neural networks have achieved state-of-the-art (SOTA) performance in almost all semantic tasks where sufficient data were available (Collobert et al., 2011; Zhang et al., 2015), neural models started showing promising results in RA only quite recently (Martine et al., 2021). A known challenge for the researchers in RA is the lack of large public datasets – with the unique exception of WeeBib (Vigliani and Meurers, 2012). Technically speaking, even WeeBib is not entirely public since it has to be directly obtained from the authors.

Martine et al. (2021) raised the SOTA classification accuracy on the popular WeeBib dataset (Vigliani and Meurers, 2012) by about 4% using BERT. This was the first solid proof that neural models with auto-generated features can show significant improvement compared to traditional ML with handcrafted features. However, neural models, or transformers (which is the interest of this paper), still show not much better performance than traditional ML on smaller datasets like OneStopEnglish (Vigliani and Lalic, 2018), despite the complexity.

From our observations, the reported low performances of transformers on small RA datasets can be accounted for two reasons: 1. Only BERT was applied to RA, and there could be other transformers that perform better, even on small datasets. 2. It is a transformer shows weak performance on small

IJCNLP

(Volume 11 Issue 8, 2020)

LXPER Index: A Curriculum-specific Text Readability Assessment Model for EFL Students in Korea

AAACL

(Workshop - Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications)

LXPER Index 2.0: Improving Text Readability Assessment for L2 English Learners in South Korea

EMNLP 2021

(Main Conference)

Pushing on Text Readability Assessment: A Transformer Meets Handcrafted Linguistic Features

*2023. 06. 03. 현재 기준

Personalized, AI-based English Education Platform Company

초격차 기술 확보를 위해 항상 연구하는 LXPER

LXPER의 자연어 처리 기술력을 입증할 수 있는
등록 완료 특허 8건, 출원 4건(해외 포함), 출원 준비 1건



특허증 유효 단어를 분류하는 장치 및 방법
특허증 영어 학습 서비스를 제공하는 서버, 방법 및 컴퓨터 프로그램
특허증 외국어 문장 빈칸 추론 문제 자동 생성 방법 및 프로그램 (x2, 분할출원 및 등록)
특허증 지문의 난이도 조절을 통하여 교육용 외국어 지문을 생성하는 방법
특허증 딥러닝 기반의 어휘 문제 자동 생성 방법
특허증 신경망 모델과 비 신경망 모델을 이용한 자연어 처리 방법 및 자연어 처리 장치
특허증 확장 및 필터 작업으로 향상된 다양한 빈칸 추론 문제 자동생성 방법 및 시스템

상태	출원일	출원 명칭	출원 구분	출원 번호
특허 출원	2020.09.23	외국어 문장 빈칸 추론 문제 자동 생성 방법 및 시스템(미국, 중국 특허출원 진행 중)	해외	PCT/KR2020/012813
특허 출원	2021.09.14	지문의 난이도 조절을 통하여 교육용 외국어 지문을 생성하는 방법, 그 컴퓨터 프로그램 및 그 장치	해외	PCT/KR2021/012529
특허 출원	2021.09.14	딥러닝 기반의 단어 의미 명확화 모델을 이용한 어휘 문제 자동 생성 방법, 그 컴퓨터 프로그램 및 그 서버 장치	해외	PCT/KR2021/012530
특허 출원	2021.11.29	LDA(Latent Dirichlet Allocation) 토픽 모델링 기법을 이용한 텍스트 분석 방법 및 이를 수행하는 텍스트 분석 장치	국내	10-2021-0167701
출원 준비중	-	지정문법교정을 통한 빈칸추론문제 선택지 자동생성 출력값 의존성 상향 기법	국내	-

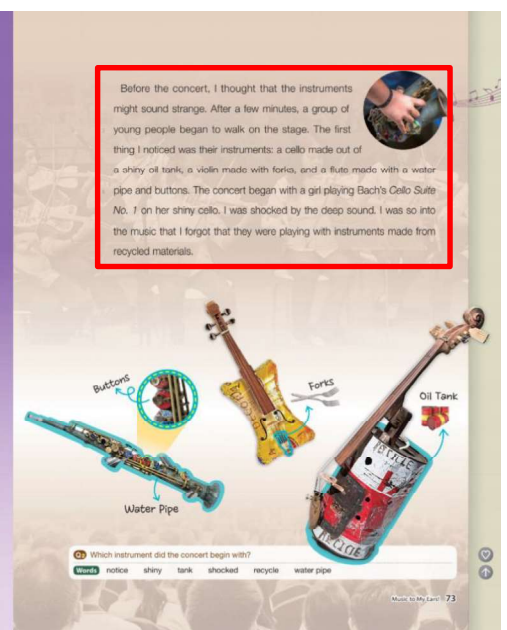
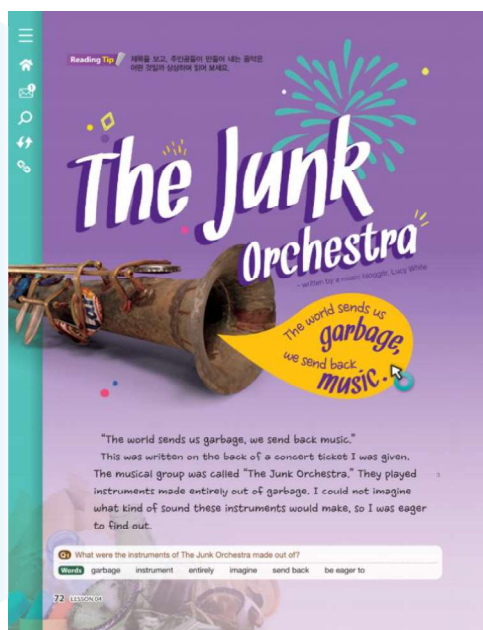
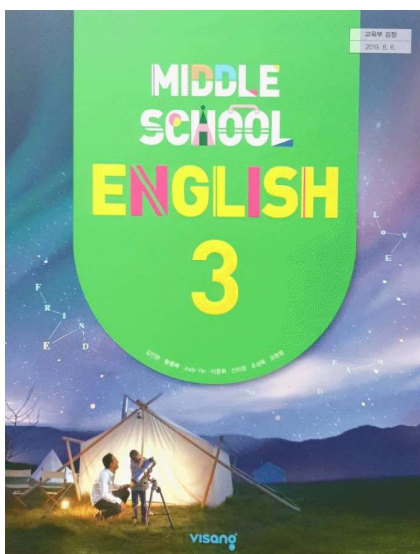
*2023. 06. 03. 현재 기준

LXPER의 기술 B2G 예시

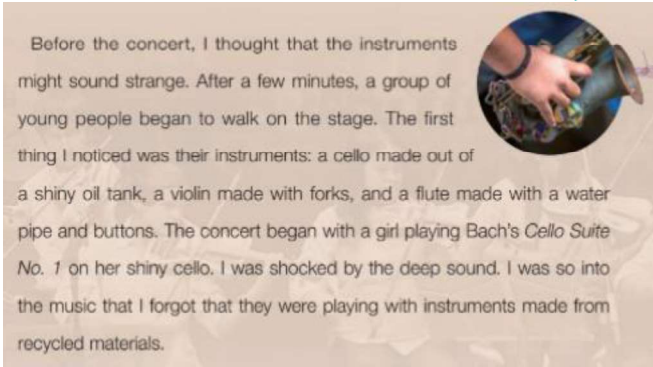


Personalized, AI-based English Education Platform Company

LXPER의 기술 B2G 예시



LXPER의 기술 B2G 예시



영어 지문을 붙여 넣으세요. 시가 난이도를 알려줘요!
보통내 정규교과과정 지문에 특화되어 있습니다.

교과과정 선택 **한국과정** **CEFR**

Before the concert, I thought that the instruments might sound strange. After a few minutes, a group of young people began to walk on the stage. The first thing I noticed was their instruments: a cello made out of a shiny oil tank, a violin made with forks, and a flute made with a water pipe and buttons. The concert began with a girl playing Bach's Cello Suite No. 1 on her shiny cello. I was shocked by the deep sound. I was so into the music that I forgot that they were playing with instruments made from recycled materials.

LXPER AI 난이도를 알려줘

LXPER Index **중3(중급)** 수준으로 예측합니다.

한국교과과정(K-12)

영어 지문을 붙여 넣으세요. 시가 난이도를 알려줘요!

교과과정 선택 **한국과정** **CEFR**

Before the concert, I thought that the instruments might sound strange. After a few minutes, a group of young people began to walk on the stage. The first thing I noticed was their instruments: a cello made out of a shiny oil tank, a violin made with forks, and a flute made with a water pipe and buttons. The concert began with a girl playing Bach's Cello Suite No. 1 on her shiny cello. I was shocked by the deep sound. I was so into the music that I forgot that they were playing with instruments made from recycled materials.

LXPER AI 난이도를 알려줘

LXPER Index **B1** 수준으로 예측합니다.

유럽언어공통기준(CEFR)

Personalized, AI-based English Education Platform Company

LXPER의 기술 B2G 예시

3 다음 글의 빈칸에 들어갈 알맞은 질문을 고르세요.

Q

A When I went to a small town called Cateura in Paraguay to work on a recycling program in 2005, I saw children living in a town that was mostly filled with garbage. I wanted to add something positive to their lives, so I decided to share my love of music with them. So I started The Junk Orchestra.

- ① Why did you visit Cateura in Paraguay in 2005?
- ② Why did you start The Junk Orchestra?
- ③ Why did you work on a recycling program?
- ④ Why did you live with the children in Cateura?

Fill in the blank question

Grammar

"Your father was not a gambler. He ㉠ sent the family money to the independence fighters in Manchuria. He made himself ㉡ look like a gambler to keep this a secret from the Japanese officers." At first, I was not sure if he was telling the truth. But afterwards, I found out the truth about my father and I realized that I ㉢ have been wrong about him. Ever since that moment, I have been proud to be the daughter of a *parakho* who ㉣ had devoted his life to the independence movement.

2 뒷글의 밑줄 친 ㉠~㉣ 중에서 어색한 것을 찾아 바르게 고치세요.

Q1 What were the instruments of The Junk Orchestra made out of?

Q2 Which instrument did the concert begin with?

Q3 What did Chávez want to share with the children of Cateura?

Q4 How did Gómez help the children to play music?

Wh- type comprehension question

Multiple choice question

1 대화를 듣고, 여자가 운영하는 블로그를 고르세요.



2 뒷글의 마지막에서 느껴지는 글쓴이의 심경으로 알맞은 것을 고르세요.

- ① sad
- ② touched
- ③ worried
- ④ bored

LXPER의 기술 B2G 예시

1. Before the concert, I thought that the instruments might sound _____.
2. After a few minutes, a group of young people began to walk on the stage. The first thing I noticed was their _____: a cello made out of a shiny oil tank, a violin made with forks, and a flute made with a water pipe and buttons.
3. The concert began with a girl playing Bach's Cello Suite No. 1 on her shiny _____.
4. I was shocked by the deep _____.
5. I was so into the music that I forgot that they were playing with instruments made from _____.

Fill in the blank question

Grammar

Which of the underlined parts of the following article is grammatically incorrect?

Before the concert, I thought that the instruments might sound ① strange. After a few minutes, a group of young people ② began to walk on the stage. The first thing I noticed ③ were their instruments: a cello made out of a shiny oil tank, a violin made with forks, and a flute made with a water pipe and buttons. The concert began with a girl ④ playing Bach's Cello Suite No. 1 on her shiny cello. I was shocked by the deep sound. I was so into the music ⑤ that I forgot that

Before the concert, I thought that the instruments might sound strange. After a few minutes, a group of young people began to walk on the stage. The first thing I noticed was their instruments: a cello made out of a shiny oil tank, a violin made with forks, and a flute made with a water pipe and buttons. The concert began with a girl playing Bach's Cello Suite No. 1 on her shiny cello. I was shocked by the deep sound. I was so into the music that I forgot that they were playing with instruments made from recycled materials.



Wh- type comprehension question

Multiple choice question

1. What materials were the instruments made from?
 - ① Metal, plastic, and wood
 - ② Oil tank, forks, and water pipe
 - ③ Glass, rubber, and stone
 - ④ Paper, cloth, and cardboard
2. How did the writer feel about the sound of the cello made from an oil tank?
 - ① Disappointed
 - ② Surprised
 - ③ Confused
 - ④ Impressed
3. What piece did the concert begin with?
 - ① Beethoven's Symphony No. 9
 - ② Chopin's Nocturne Op. 9 No. 2
 - ③ Bach's Cello Suite No. 1
 - ④ Vivaldi's Four Seasons
4. How did the writer react to the music?
 - ① They felt bored

Personalized, AI-based English Education Platform Company

LXPER의 기술 B2G 예시

영어 지문을 붙여 넣으세요. AI가 난이도를 알려줘요!

* 국내 정규교과과정 지문에 특화되어 있습니다.

교과과정 선택 한국국문 CEFR

Before the concert, I thought that the instruments might sound strange. After a few minutes, a group of young people began to walk on the stage. The first thing I noticed was their instruments: a cello made out of a shiny oil tank, a violin made with forks, and a flute made with a water pipe and buttons. The concert began with a girl playing Bach's Cello Suite No. 1 on her shiny cello. I was shocked by the deep sound. I was so into the music that I forgot that they were playing with instruments made from recycled materials.

난이도: 100, 분량: 97 /

LXPER AI, 난이도를 알려줘

LXPER Index: **중3(중급)** 수준으로 예측합니다.

원본 지문(중3_중급)

영어 지문을 붙여 넣으세요. AI가 난이도를 알려줘요!

* 국내 정규교과과정 지문에 특화되어 있습니다.

교과과정 선택 한국국문 CEFR

I thought the instruments would sound weird before the concert. But then, some young people played music on unique instruments made from recycled things like an oil tank, forks, and a water pipe. They played Bach's Cello Suite No. 1 and I liked the deep sound. I forgot about the strange instruments and just listened to the music.

난이도: 58, 분량: 5 /

LXPER AI, 난이도를 알려줘

LXPER Index: **중2(고급)** 수준으로 예측합니다.

더욱 쉬운 지문으로(중2_고급)

더욱 어려운 지문으로(고3_고급)

Prior to the commencement of the performance, I harbored concerns regarding the potential peculiarity of the instruments to be utilized. However, as a troupe of young individuals strode onto the stage, my focus swiftly gravitated towards their visually striking instruments - a cello fabricated from a polished oil reservoir, a violin fashioned from forks, and a flute crafted using a water pipe adorned with buttons. The concert was launched with a rendition of Bach's Cello Suite No. 1, being beautifully executed by a young lady playing the aforementioned cello. The profundity of the melody rendered me stupefied, to the point where I had completely forgotten that these were indeed, recycled instruments.

난이도: 110, 분량: 5 /

LXPER AI, 난이도를 알려줘

LXPER Index: **고3(고급)** 수준으로 예측합니다.



Personalized, AI-based English Education Platform Company

AI기반 1:1 개인 맞춤 교육

22



AI기반 1:1 개인 맞춤 교육

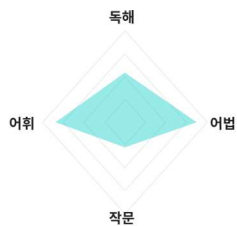
학생 정보



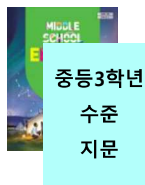
학생 A

- 초등학교 6학년
- 외국 체류 경험 있음

영역별 수준



오프라인 수업



AI 기반 온라인 과제



- 쉬운 독해/문법 과제
- 어려운 어휘/작문 과제

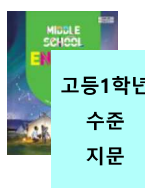
실력 예측

“학생은 A중학교2학년1학기 중간고사에서 90% 확률로 93점을 획득할 것으로 예측됩니다.”



학생 B

- 중학교 3학년
- 외국 체류 경험 없음



- 어려운 독해/문법 과제
- 쉬운 어휘/작문 과제

“학생은 이번 중학교 중간고사에서 80% 확률로 90점을 획득할 것으로 예측됩니다.”

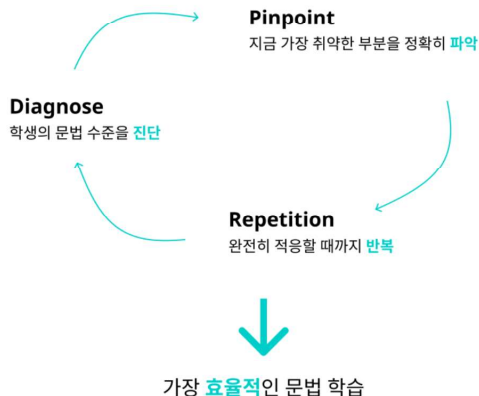
Personalized, AI-based English Education Platform Company

AI기반 1:1 개인 맞춤 교육의 시작. Grammar PEEP

AI 영문법 관리 시스템 **그래머핍**

Grammar PEEP

[Grammar Predictor & Evaluator for English Proficiency]



선생님, 영어 문법 수업 뒤에 학생들이 모두 잘 이해했는지 걱정되시나요?

아이들의 문법 실력이 너무 제각각이에요.


많은 질문에 모두 답변하기엔 시간이 부족해요.

어떤 문법 교재를 선택하면 좋을지 모르겠어요.

학생이 문법 때문에 영어 자체가 무섭고, 어렵다고 생각해요.

학부로 상담, 참고할만한 자료가 필요해요.

AI기반 1:1 개인 맞춤 교육의 시작. Grammar PEEP



문장이란?

	BASIC LEVEL 1	BASIC LEVEL 2	BASIC LEVEL 3
문장이란?			
동사			
명사			
대명사			
형용사			
부사			
조동사			
의문사			
관사			
접속사			
관사			

과제 리포트

45 점 **종결종 - be 동사 과거형**

학습 일시 - 23.2.6

학습 시간 - 13:25

학습 점수 - 97 점

틀어 놓음 - 총 32개 / 정답 30개 / 오답 2개

정답률 - 93.8%

GRAMMAR PEEP'S Comment

종결종 학생은 be 동사 과거형의 개념을 완벽하게 이해하고, 대부분의 문제를 잘 풀었습니다. 다만, 고난도 문제를 푸는 연습을 더 하면 좋을 것 같습니다.

종결종 학생 틀어 문항 확인하기

(업데이트 예정입니다.)

Grammar PEEP

안녕하세요, Grammar PEEP입니다.

종결종 학생의 **be 동사 과거형** 학습 리포트입니다.

학습 일시 - 2023. 2. 6

학습 시간 - 13:25

학습 점수 - 97점

틀어 놓음 - 총 32개 / 정답 30개 / 오답 2개

정답률 - 93.8%

Grammar PEEP's Comment

종결종 학생은 be 동사 과거형의 개념을 완벽하게 이해하고, 대부분의 문제를 잘 풀었습니다. 다만, 고난도 문제를 푸는 연습을 더 하면 좋을 것 같습니다.

상세보기

4. Match the contractions to the words.

i'll aren't

it is it's

are not

2. Put the sentences in order on the timeline.

Cole sits by the window in class.

Cole will sit with Jackson at lunch.

Cole sat next to me on the bus.

past present (now) future

3. Type the contraction that matches the words.

she will = 정답을 입력해주세요. 띄어쓰기가 틀린 경우 오답 처리됩니다.

Personalized, AI-based English Education Platform Company

ADB와 함께하는 글로벌 비즈니스



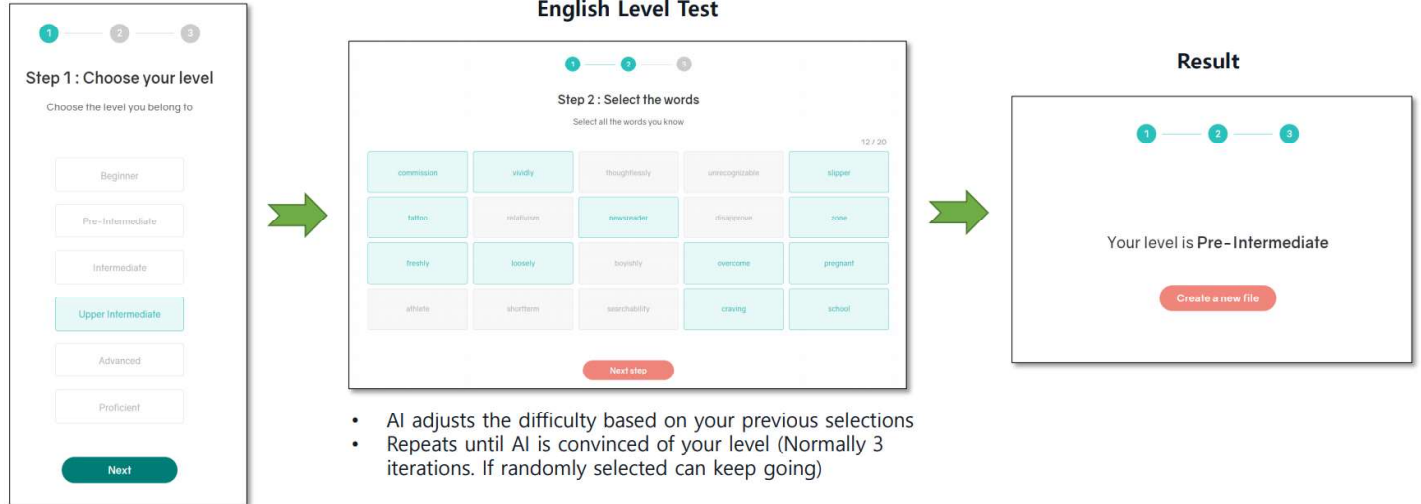
AI Application for Primary Education in Bangladesh

ByeongJo Kong
 Digital Technology Specialist (Data Analytics and Big Data)
 Digital Technology for Development Unit,
 Sustainable Development and Climate Change Department
 Asian Development Bank (ADB)

AI Solution – Automated AI Test Maker

1. English Level Assessment

Selection of Difficulty Level



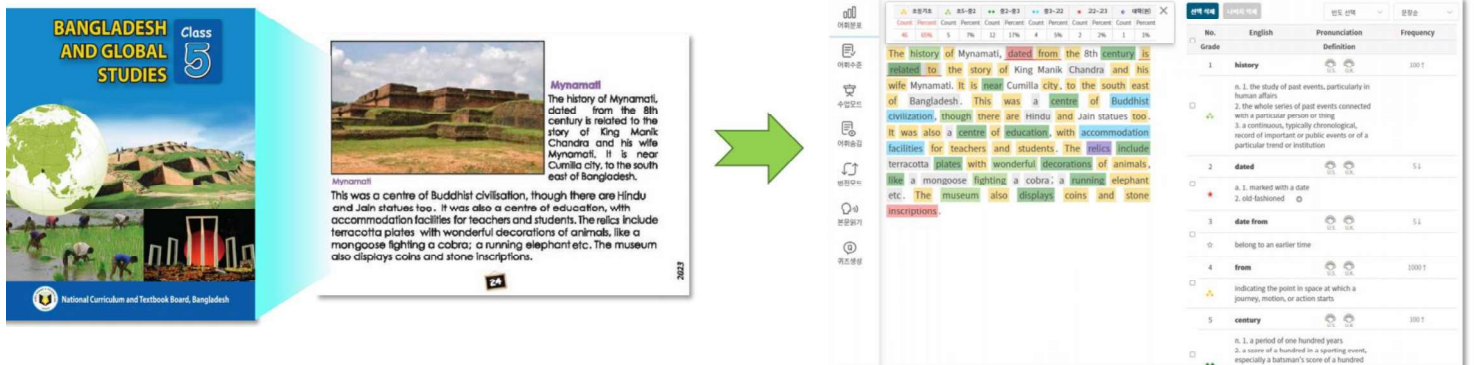
*ADB의 방글라데시 교육부 대상 실제 발표 자료 발췌

Personalized, AI-based English Education Platform Company

AI Solution 1 – Automated AI Test Maker

2. Personalize Learning Contents

Vocab-Level Text Analysis and Vocab Classification by Difficulty Levels



*ADB의 방글라데시 교육부 대상 실제 발표 자료 발췌

Personalized, AI-based English Education Platform Company

AI Solution 1 – Automated AI Test Maker

2. Personalize Learning Contents

Personalize Difficulty Levels

This engine is optimized for the Korean standard curriculum.

Select Curriculum: **Korea** **CEFR**

The history of Mynamati, dated from the 8th century is related to the story of King Manik Chandra and his wife Mynamati. It is near Cumilla city, to the south east of Bangladesh. This was a centre of Buddhist civilization, though there are Hindu and Jain statues too. It was also a centre of education, with accommodation facilities for teachers and students. The relics include terracotta plates with wonderful decorations of animals, like a mongoose fighting a cobra; a running elephant etc. The museum also displays coins and stone inscriptions.

Word Count: 90, Sentence Count: 6

CALCULATE READABILITY

Index: **MID-3 (Intermediate) Level**

Make harder
(from MID-3 to HIGH-2)

Make easier
(from MID-3 to MID-2)

This engine is optimized for the Korean standard curriculum.

Select Curriculum: **Korea** **CEFR**

Mynamati, situated to the southeast of Bangladesh's Cumilla city, is steeped in history, dating back to the 8th century. This revered site is inextricably linked to the tale of King Manik Chandra and his wife Mynamati. The center of Buddhist civilization, Mynamati also features Hindu and Jain statues. The locale was a hub of education and boasted accommodation facilities for teachers and students alike. A visit to Mynamati affords an opportunity to behold an array of remarkable relics, including terracotta plates, intricately adorned with depictions of animals such as a mongoose locked in combat with a cobra and a galloping elephant. Moreover, the museum showcases a variety of ancient coins and stone inscriptions, offering valuable insights into the region's past.

Word Count: 120, Sentence Count: 6

CALCULATE READABILITY

Index: **HIGH-2 (Advanced) Level**

This engine is optimized for the Korean standard curriculum.

Select Curriculum: **Korea** **CEFR**

Mynamati is a historical place in Bangladesh. It has a lot of old things that are from a long time ago. People used to learn there and it was an important place for education. There are many things to see, such as pictures of animals like a mongoose fighting a cobra, and an elephant running. There are also old coins and writings on rocks that you can see. Mynamati is a special place that has many interesting things to explore!

Word Count: 80, Sentence Count: 6

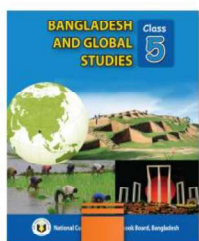
CALCULATE READABILITY

Index: **MID-2 (Beginner) Level**

- Analyzes the article-level difficulty level
- Bangladesh difficulty level analysis need to be added

AI Solution – Automated AI Test Maker

3. Assessment Question Creation



This engine is optimized for the Korean standard curriculum.

Select Curriculum: **Korea** **CEFR**

The history of Mynamati, dated from the 8th century is related to the story of King Manik Chandra and his wife Mynamati. It is near Cumilla city, to the south east of Bangladesh. This was a centre of Buddhist civilization, though there are Hindu and Jain statues too. It was also a centre of education, with accommodation facilities for teachers and students. The relics include terracotta plates with wonderful decorations of animals, like a mongoose fighting a cobra; a running elephant etc. The museum also displays coins and stone inscriptions.

Word Count: 90, Sentence Count: 6

CALCULATE READABILITY

Index: **MID-3 (Intermediate) Level**

Enter texts and select question types

The ²h_____ of Mynamati, dated from the 8th century is ¹_____ to the story of King Manik Chandra and his wife Mynamati. It is near Cumilla city, to the south east of Bangladesh. This was a centre of Buddhist civilization, though there are Hindu and Jain statues too. It was also a centre of education, with accommodation facilities for teachers and students. The relics include terracotta plates with wonderful decorations of animals, like a mongoose fighting a cobra; a running elephant etc.

<호 기>
history, include

The ¹_____ of Mynamati, dated from the 8th century is related to the story of King Manik Chandra and his wife Mynamati. The relics ²_____ with wonderful decorations of animals, like a mongoose fighting a cobra; etc.

Fill in the blank question

True or False question

The history of Mynamati, dated from the 8th century is related to the Chandra and his wife Mynamati. It is near Cumilla city, to the south
This was a centre of Buddhist civilization, though there are Hindu and Jain statues too. It was also a centre of education, with accommodation facilities for teachers and students. The relics include terracotta plates with wonderful decorations of animals, like a mongoose fighting a cobra; a running elephant etc. The museum also displays coins and stone inscriptions.

- 1) The history of mynamati dates back to the 8th century. (T / F)
- 2) King manik chandra's story about mynamati is from the 17th century. (T / F)
- 3) East of bangladesh is close to cumilla city. (T / F)
- 4) It is to the north-west of bangladesh. (T / F)
- 5) There are statues of both hindu and jain in mynamati. (T / F)

The history of Mynamati, ⁴dating from the 8th century is related to Manik Chandra and his wife Mynamati.

1) _____ dating _____ → _____

This was a centre of Buddhist civilization, though there ⁵is Hindu and

2) _____ is _____ → _____

The relics include terracotta plates with wonderful decorations of animals, like a mongoose fighting a cobra; a running elephant etc.

⁶fought a cobra; a running elephant etc.

3) _____ fought _____ → _____

Grammar questions

Multiple choice questions

_____ plates can be seen at the Mynamati museum.
a) Ceramic plates
b) Terracotta plates
c) Porcelain plates
d) Terracotta plates

2. What type of civilization was centered in Mynamati?
a) Christian b) Muslim
c) Buddhist d) Hindu

3. What type of facilities were available for education at Mynamati?
a) Sports facilities b) Science labs
c) Accommodation facilities d) Cafeterias

4. What is the story of King Manik Chandra related to Mynamati?
a) The founding of the city
b) The history of the royal family

Address : 3F, 14-2, Apgujeong-ro-4-gil, Gangnam-gu, Seoul, Korea (06027)

Tel : 82-2-6338-7905

E-mail : jasonlee@lxper.com (Founder&CEO)



In The New Normal Era

Global Edtech Frontier, LXPER

www.lxper.ai

On Repair of the Subject Condition by Parasitic Gaps*

Sun-woong Kim (Kwangwoon University, Seoul)

2023/6/3

한국코퍼스언어학회-한국영어학회 2023년 여름 공동 학술대회, 고려대학교

*This works was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2022S1A5A2A0103824711)

Setting Stage: Subject Condition Violation and Repair by Parasitic Gaps

- (4) a. *a person who [(not) shaking hands with t] would really bother Sally
 - b. *a person who [us shaking hands with t] would really bother Sally
 - c. *a person who [Terry shaking hands with t] would really bother Sally

 - (5) a. a person who [(not) shaking hands with e] would bother t
 - b. a person who [us shaking hands with e] would bother t
 - c. a person who [Terry shaking hands with e] would bother t
- (e = parasitic gap; t = real gap)

1. Introduction

- The elements inside the subject cannot be extracted, which is known as the Subject Condition (Chomsky 1982). From this perspective, look at the following complex-looking examples (Culicover and Winkler 2022, CW)
- (1) a. *a person who_i [talking to t_i] about this would prove to the Mayor that there is a problem.
b. a person who_i [talking to e_i] about this would prove to t_i that there is a problem
(e = parasitic gap; t = real gap)

3

- (2) {who, t, the Mayor} vs.
{who, e, t}
(e = parasitic gap; t = real gap)
- the Mayor = **uninvited guest**

4

2. Extraction from the Subject and Parasitic Gaps

Major Claims of CW

- (3) a. There is no specific parasitic gap mechanism that has the effect of making extraction from a subject island grammatical.
- b. The link between the two is a matter of processing complexity, and that the amelioration associated with the parasitic gap configuration is a consequence of differences in complexity.

5

- They further elaborate that the extraction out of the subject is itself acceptable, marginally complex though.
- The introduction of a new discourse referent, however, "increases complexity and pushes the sentence over the threshold of unacceptability", eventually leading to the judgment of 'ungrammaticality'.
- If stated from the opposite point of view, the absence of a referring expression (i.e. the uninvited guest) keeps the sentence closer to an acceptable status.
- All in all, after rejecting what they call "Syntactic Islands Hypothesis" and "The Gap Repair Hypothesis", they support "**The Uninvited Guest Hypothesis**" for the contrast given in (1).

6

2.1 Subject Condition and Repair

- The typical examples of the Subject Condition are given below. All the examples are ungrammatical in that some element is extracted out of the subject.

- (4)
- a. *a person who [(not) shaking hands with t] would really bother Sally
 - b. *a person who [us shaking hands with t] would really bother Sally
 - c. *a person who [Terry shaking hands with t] would really bother Sally

7

- In contrast, if the referring direct object in (4) is replaced by a true gap and the gap in the subject is replaced by a parasitic gap, grammaticality is dramatically improved.

- (5)
- a. a person who (not) shaking hands with e would bother t
 - b. a person who us shaking hands with e would bother t
 - c. a person who Terry shaking hands with e would bother t

8

- The repair by parasitic gaps is possible in adjunction but is not in coordination (Chaves 2012, Overfelt 2015).
- (6) a. *Sam interviewed *ti* and showed *ti* his secret laboratory, [the members of the incoming class of graduate students]*i*.
 b. Sam interviewed *ti* before showing *ei* his secret laboratory, [the members of the incoming class of graduate students]*i*.
- (7) a. *Kim surprised *ti* and offered *ti* a raise, [everyone who showed up early]*i*.
 b. Kim surprised *ti* by offering *ei* a raise, [everyone who showed up early]*i*.

9

2.2 Invited Guests

- (13) a. ... spirit fond of encountering difficulties, which to avoid would incur no censure, but which to meet would be impertinent and hazardous.
- b. By this she meant, if someone is making me feel bad about myself or is causing me pain and upset, they should be in my life, and that I should value those who spending time with always feels like a positive experience.

10

2.3 Challenges to Uninvited Guests

(15) modified from Chaves (2012)

- a. That is the photo that I found t without looking for e.
- b. [?]That is the photo that I found Mom in without looking for t.
- c. [?]That is the photo that I found t while looking for e.

11

- (16)
- a. That is the library that I hung around in t without looking for e.
 - b. ?That is the library where I spotted the book without looking for t.
 - c. That is the library that I hung around in t while looking for a book.

12

2.4 Uninvited Guest Issues in Korean

- (20) a. [__ Mary-lul cafe-esey manna-n] namca
M-ACC cafe-at meet-rel man (rel=relative pronoun)
'a man that met Mary at the cafe'
- b. [John-i __ cafe-esey manna-n] yeca
J-NOM cafe-at meet-rel woman
'a woman that John met Mary at the cafe'
- c. [John-i Mary-lul __ manna-n] cafe
J-NOM M-ACC meet-rel cafe
'a cafe where John met Mary'

13

- (21) a. [[__ Mary-lul cafe-esey mannasstako] ney-ka sayngkakhanu-n] namca
M-ACC cafe-at meet-rel you-NOM think-rel man
'a man who you that met Mary at the cafe'
- b. [[John-i __ cafe-esey mannasstako] ney-ka sayngkakhanu-n] yeca
J-NOM cafe-at meet-rel you-NOM think-rel woman
'a woman who you think that John met Mary at the cafe'

14

- (22) a. ?[[pro t hug-lul hamyen] [John-lul koylophikye toynu-n] namca
 hug-ACC do-if J-ACC tease become-rel man
 'a man who if pro hugs will tease Mary'
- b. ??[[pro e hug-lul hamyen] [t koylophikye toynu-n] namca
 hug-ACC do-if tease become-rel man
 'a man who if pro hugs will be teased'

15

3. Proposal

3.1. Proposed Analysis

- (23) a. This problem was to have been looked into, but obviously, nobody did.
 b. Roses should have been brought by somebody, but surprisingly, nobody did.
- (24) a. *This problem was looked into by John, and (similarly) Bob did, too.
 b. *Roses were brought by someone, and others did.

16

(25) (=8)

- a. *Sam interviewed t *and* showed t his secret lab, [the members of ...].
- b. Sam interviewed t *before* showing e his secret lab, [the members of ...].

(26) *John [_{Transitive} saw t] and Bill [_{Ditransitive} gave t a book], Mary.

17

- (27) a. *a person who_i [talking to t_i] about this would prove to the Mayor that there is a problem.
- b. a person who_i [talking to e_i] about this would prove to t_i that there is a problem (e = parasitic gap; t = real gap)

- (28) a. ... who_i [_{SubjP} PRO talking to t_i] ...
- b. ... who_i [_{PragP} [_{SubjP} PRO talking to e_{pgi}]] ...

18

3.2. Consequences

3.2.1. LBC

(30) Quantity *how*

- a. i. *How many should I buy ___ peppers for the dinner?
ii. *How much did she find ___ gold?
- b. i. I should buy some peppers for the dinner, but I don't know how many.
ii. She found gold, but won't say how much.

(31) Degree *how*

- a. *How does she want [a ___ detailed list]?
- b. i. *He wants a detailed list, but I don't know how.
ii. *She bought an expensive car, but I don't know how.

19

3.2.1. Adjunct Islands

(33) a. a person who Susan went to Washington to work with t

b. *War and Peace* is a book that I always fall asleep while reading t.

(34) **War and Peace* is a book that I always amuse Sandy [_{AdjunctP} while reading to work with t].

(35) *War and Peace* is a book that I always fall asleep with t [_{PragP} [_{AdjunctP} while reading eg].

20

4. Conclusion

- This paper showed that the uninvited guest hypothesis has both over and undergeneration problems.
- To overcome the difficulties, this paper proposed that the RPG effect is real and that syntactic analysis is more plausible from derivational perspectives.
- In particular, this paper that the derivationally dynamic phases can provide a better solution to the effect that covers the examples with strong islands like LBC and that-t effect.
- Eventually if the conclusions of this paper are correct, this paper gives tribute to the effort to incorporate discourse-pragmatic notions into syntax, leading to a collaborative syntax, which is an evolved stage of autonomous syntax.

Homonyms-based word replacement algorithm for generating adversarial examples

유 은 영 (충남대학교)

강 아 름 (충남대학교)

1. Introduction

1.1. Motivation

- Replacing words has been used to extend corpus data.
- Kim Jae-kyun(2021) augmented the Korean named entity tagged corpus by using synonyms, and reported that the performance improved when an augmented corpus was added to the existing corpus.
- In a similar context, it would be possible to expand the existing corpus using homonyms and their synonyms.

Quesetion: Is it possible to fool the model using homonyms-based word replacement algorithm for generating adversarial examples?

- Goal

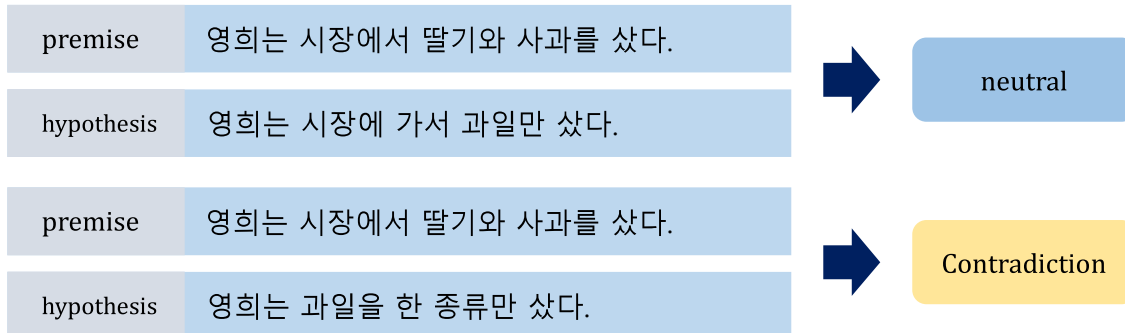
To examine how much homonyms-based word replacement algorithm can fool the model in the course of natural language inference machine learning

2. Previous Studies

2.1. Analysis of cross-text implication relationship

- Proper **labeling** of a given pair of sentences (premise sentences & hypotheses) by **inferring** whether the relationship between the two sentences is **entailment, neutral, or contradictory**
- Entailment: definitely correct
- Neutral: neither definitely correct nor definitely incorrect
- Contradiction: definitely incorrect

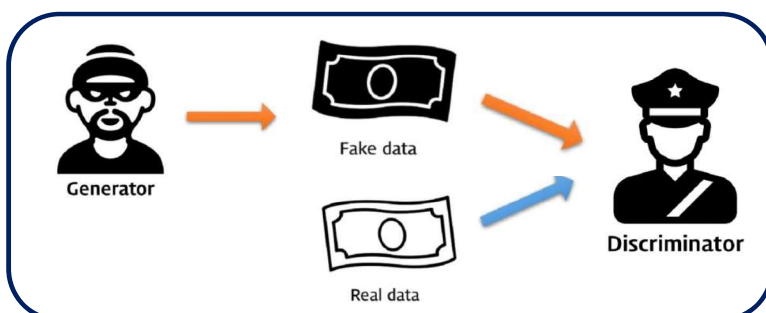
2022 Adversarial Implications Analysis Explanatory Material



2. Previous Studies

2.2. Adversarial Examples

- Data that **intentionally mixes noise** with the original context, which aims to fool the natural language processing model that has received the adversarial examples as learning data
- For example, the original implication relationship of an adversarial example is contradictory, but inputting the example in the model leads to misinterpretation of the implication relationship as entailment or neutral. (2022 Adversarial Implications Analysis Explanatory Material)
- Inputting adversarial examples as learning data on several natural language processing models, including BERT, has shown to improve robustness of NLP models. (Jin et al., 2020)

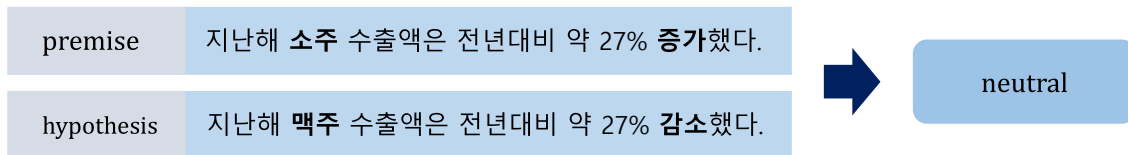


생성적 적대 신경망 소개[Website]. (2023, May 27).
https://paseul.github.io/deeplearning/gan_introduction/

2. Previous Studies

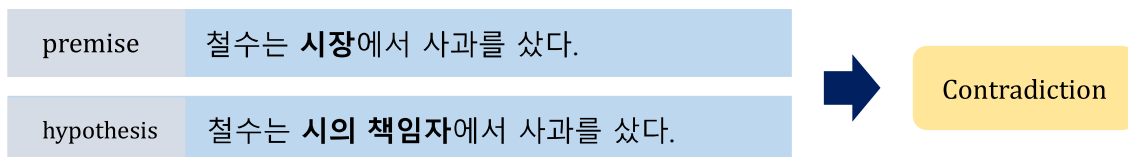
2.3. Lexical

- One of the **inference methods** that can generate adversarial examples
- **Lexical information** about things like synonyms and antonyms



2022 Adversarial Implications Analysis Explanatory Material

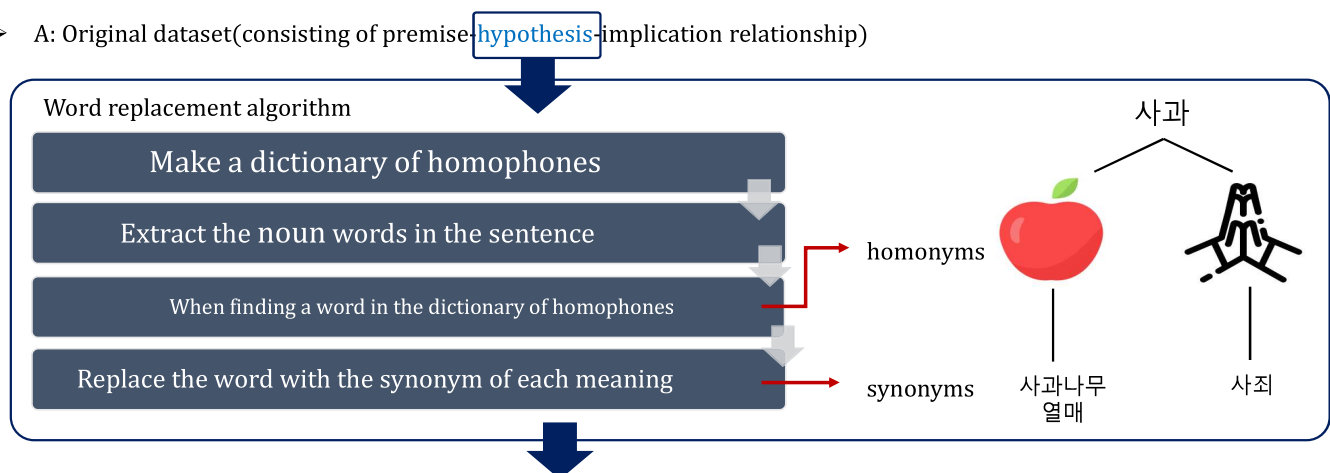
- An example of using homonym information to make adversarial cases as part of a lexical inference method



- The Korean word '시장' is a homophone with two meanings, 'market' and 'head of the city'.
- In the sentence, the original word '시장' was changed to the meaning corresponding to the side that does not fit the context, so that the relationship of implication was contradiction.

3. Research Method

- Dataset: KLUE NLI Dataset(<https://klue-benchmark.com/tasks/68/data/description>)
- Model: Roberta
- Outline
 - A: Original dataset(consisting of premise-hypothesis-implication relationship)



- B: Expanded dataset(consisting of premise-new hypothesis-implication relationship)
- Train the model on A and B, and find the accuracy of the result.

3. Research Method

3.2. Procedure

- Make a dictionary of homophones
- Korean Vocabulary Learning List from the National Institute of Korean Language(https://www.korean.go.kr/front/etcData/etcDataView.do?mn_id=46&etc_seq=71)

단어	품사	풀이
1195 가게	명	
898 가격03	명	價格
2986 가구03	명	家口
7434 가구04	명	家具
4964 가까워지다	동	
1695 가까이	부	
3560 가까이	명	
648 가깝다	형	
1747 가꾸다	동	
852 가끔	부	
5071 가난01	명	
1716 가난하다	형	
2757 가늘다	형	
6942 가능	명	可能
483 가능성	명	可能性
501 가능하다	형	可能-
6003 가능해지다	동	可能-
25 가다01	동	
150 가다01	보	
1196 가득01	부	~ 차다
3561 가득하다01	형	

밤01	명	어두운 때
밤02	명	열매

가구03	명	家口
가구04	명	家具

사과05	명	沙果
사과08	명	謝過

Words that have homonyms are given a superscript number in the dictionary.
-> Extract nouns with superscript number to make a dictionary of homophones

3. Research Method

3.2. Procedure

- Make a dictionary of homophones

```

korean_db = pd.read_excel('C:\\Users\\user\\spyder-py3\\한국어학습용어휘목록.xlsx')
numbers = "0123456789"
conditions = (korean_db["단어"].str.contains(f"[{numbers}]")) & (korean_db["품사"] == "명")
homonym_db = korean_db.loc[conditions]

homonym_db = homonym_db.replace({"단어": "[0123456789]"}, {"단어": ""}, regex=True)

noun_dict = {}
for noun in noun_list:
    if noun in homonym_db["단어"].values:
        pos = homonym_db.loc[homonym_db["단어"] == noun, "품사"].values[0]
        if pos == "명":
            noun_dict[noun] = homonym_db.loc[homonym_db["단어"] == noun, "풀이"].values[0]

duplicated_nouns = []
for noun in noun_dict.keys():
    if list(noun_dict.keys()).count(noun) > 1:
        duplicated_nouns.append(noun)
        print(f"{noun}: {noun_dict[noun]}")

homonym_db.to_excel('C:\\Users\\user\\Desktop\\outputs.xlsx', index=False)
    
```

Initial state of column D

풀이	B	C	D
價格	거리	명	간격
家口	검사	명	검찰관
家具	검사	명	체크
	경기	명	호황과 불
	경기	명	겨루기
	고개	명	목덜미
	고개	명	연넝
	고장	명	지역
	고장	명	기능상의
	공기	명	대기
	공기	명	밥공기
	과	명	학과
	과	명	업무 부서
	과거	명	과거시험
可能	과거	명	지나간 때
可能性	과정	명	단계
可能-	과정	명	절차
可能-	관리	명	관료
	관리	명	간수
	군	명	군대
~ 차다	군	명	군 소재지
	군사	명	군인
	군사	명	군무

- Each of the '풀이' columns written in Chinese characters is modified with words that can be replaced in the sentence.

3. Research Method

3.2. Procedure

- Homonyms-based word replacement algorithm

- Download KLUE NLI dataset

guid	source	premise	hypothesis	label
klue-nli-v1_t_NSMC		햇걸 진심 최고다 그 어떤 히어로보다 멋지다	햇걸 진심 최고로 멋지다.	0
klue-nli-v1_t_NSMC		100분간 잘겔 그레도 소닉붐뎀에 2점준다	100분간 잤다.	2
klue-nli-v1_t_NSMC		100분간 잘겔 그레도 소닉붐뎀에 2점준다	소닉붐이 정말 멋있었다.	1
klue-nli-v1_t_NSMC		100분간 잘겔 그레도 소닉붐뎀에 2점준다	100분간 자는게 더 나았을 것	1
klue-nli-v1_t_airbnb		101빌딩 근처에 나름 즐깁거리가 많습니다.	101빌딩 근처에서 즐깁거리 찾	2
klue-nli-v1_t_airbnb		101빌딩 근처에 나름 즐깁거리가 많습니다.	101빌딩 주변에 젊은이들이 즐	1
klue-nli-v1_t_airbnb		101빌딩 근처에 나름 즐깁거리가 많습니다.	101빌딩 부근에서는 여러가지	0
klue-nli-v1_t_wikipedia		10년 만에 찾는 피터를 웬디는 따뜻하게 맞이하고	1 웬디는 피터를 차갑게 맞이했다	2
klue-nli-v1_t_wikipedia		10년 만에 찾는 피터를 웬디는 따뜻하게 맞이하고	1 잭과 매기는 피터 배닝의 동생	1
klue-nli-v1_t_wikipedia		10년 만에 찾는 피터를 웬디는 따뜻하게 맞이하고	1 피터 배닝, 잭, 매기는 남매사이	0
klue-nli-v1_t_wikipedia		10년 전 한 병원에서 입원 중인 한 소녀가 실종되는 병원에서 실종된 소녀는 경찰들		2
klue-nli-v1_t_wikipedia		10년 전 한 병원에서 입원 중인 한 소녀가 실종되는	10년 전 한 병원에서 입원 중인	0
klue-nli-v1_t_wikipedia		10년 전 한 병원에서 입원 중인 한 소녀가 실종되는	10년 후 과학 기술이 발달되어	1
klue-nli-v1_t_wikipedia		10년 전, 공동육아를 매개로 성미산 마을에 들어와	강석필, 홍형숙 부부가 성미산	2

Labels represent implication relationships.
 0: entailment
 1: neutral
 2: contradiction

The goal is to increase the number of adversarial examples by **replacing words in hypothesis sentences**.

- After tokenizing hypothesis sentences, if there are noun words in the homonym dictionary, replace the noun words with words in the dictionary's '풀이' column.

3. Research Method

3.2. Procedure

- Homonyms-based word replacement algorithm

- Select 400 random sentences

- Get new hypothesis sentences for the original hypothesis sentences by running the algorithm

```
def make_new_sentence(sentence):
    nouns = tokenizer.nouns(sentence)
    synonyms = []
    new_sentence_list = []
    for noun in nouns:
        synonyms = []
        for i in range(len(df2)):
            if noun == df2.loc[i, "단어"]:
                synonyms.extend(df2.loc[i, "풀이"].split(","))
        if len(synonyms) == 0:
            continue
        for synonym in synonyms:
            new_sentence = sentence.replace(noun, synonym)
            new_sentence_list.append(new_sentence)
    return new_sentence_list
```

```
df2 = pd.read_excel("data.xlsx")

result_list = []
for i in range(len(df2)):
    premise = df2.loc[i, "premise"]
    hypothesis = df2.loc[i, "hypothesis"]
    new_sentences = make_new_sentence(hypothesis)
    for new_sentence in new_sentences:
        result_list.append([premise, new_sentence])

result_df2 = pd.DataFrame(result_list, columns=['premise', 'new_hypothesis'])
result_df2.drop_duplicates(inplace=True)
```



hypothesis	label
경주에서 오래된 시장은 시내 한복판에 자리하고 있다.	0
신문증 위조에 사용하기 어렵도록 하는 기능이 있다.	0
송영호 씨와 같은 정기 헌혈자의 수는 점차 감소하고 있다.	1
약 100명이 신청했다.	1
한글의 역사에 대해 생생하게 전할 수 있게 해주는 유물들도	0
영화도 끝까지 봐야 그 진가를 알 수 있다.	0

new_hypothesis
경주에서 오래된 마켓은 시내 한복판에 자리하고 있다.
경주에서 오래된 시의 책임자는 시내 한복판에 자리하고 있다.
경주에서 오래된 시장은 시내 한복판에 공간하고 있다.
경주에서 오래된 시장은 시내 한복판에 돛자리하고 있다.

3. Research Method

3.2. Procedure

- Homonyms-based word replacement algorithm

➤ As a result, we get about 1000 sentence pairs.

premise	new_hypothesis	label
시내랑은 걸어서	걸어서 시내에 갈 방법 있습니다.	0
시내랑은 걸어서	걸어서 시내에 갈 숫자 있습니다.	1
또한 스크린 캡처	신분증 위조에 사용하기 어렵도록 하는 기량이 있다.	0
또한 스크린 캡처	신분증 위조에 사용하기 어렵도록 하는 쓰임이 있다.	0

- Enter the correct answer label for accuracy comparison after learning the model.
- Train the model and input the sentences.
- Compare the accuracy of training on 400 original sentence pairs with the accuracy of training 1000 extended sentence pairs.

3. Research Method

3.2. Procedure

- Train the model

1) Install the libraries

```
!pip install --upgrade accelerate
!pip install transformers
!pip install datasets
!pip install scipy
!pip install scikit-learn

import datasets
import logging
import numpy as np
import pandas as pd
import random
from IPython.display import display, HTML
from datasets import load_dataset, load_metric, ClassLabel, Sequence
from transformers import AutoTokenizer, AutoModelForSequenceClassification, TrainingArguments, Trainer
```

2) set up the options for model

```
task = "nli"
model_checkpoint = "klue/roberta-base"
batch_size = 32
```

3) Download the KLUE-NLI Dataset

Set up the metric for evaluation

Load the tokenizer

Preprocess the dataset

```
datasets = load_dataset("klue", task)
datasets["train"][0]
metric = load_metric("glue", "qnli")

# generate the random datasets
fake_preds = np.random.randint(0, 2, size=(64,))
fake_labels = np.random.randint(0, 2, size=(64,))
fake_preds, fake_labels

metric.compute(predictions=fake_preds, references=fake_labels)
tokenizer = AutoTokenizer.from_pretrained(model_checkpoint, use_fast=True)

# set up the premise and the hypothesis
sentence1_key, sentence2_key = ("premise", "hypothesis")
print(f"Sentence 1: {datasets['train'][0][sentence1_key]}")
print(f"Sentence 2: {datasets['train'][0][sentence2_key]}")

def preprocess_function(examples):
    return tokenizer(
        examples[sentence1_key],
        examples[sentence2_key],
        truncation=True,
        return_token_type_ids=False,
    )

preprocess_function(datasets["train"][:5])
encoded_datasets = datasets.map(preprocess_function, batched=True)
```


3. Research Method

3.2. Procedure

- Train the model

4) Define the Model

```
num_labels = 3

# define the model
model =
AutoModelForSequenceClassification.from
_pretrained(model_checkpoint,
num_labels=num_labels)
```

5) Train the model

```
# define the function for computing metrics
def compute_metrics(eval_pred):
    predictions, labels = eval_pred
    predictions = np.argmax(predictions, axis=1)
    return metric.compute(predictions=predictions, references=labels)

# define the metric name
metric_name = "accuracy"

!pip install --upgrade transformers
from transformers import TrainingArguments

args = TrainingArguments(
    "test-nli",
    evaluation_strategy="steps",
    learning_rate=2e-5,
    per_device_train_batch_size=batch_size,
    per_device_eval_batch_size=batch_size,
    num_train_epochs=5,
    weight_decay=0.01,
    load_best_model_at_end=True,
    metric_for_best_model=metric_name,
)

# define the trainer
trainer = Trainer(
    model,
    args,
    train_dataset=encoded_datasets["train"],
    eval_dataset=encoded_datasets["validation"],
    tokenizer=tokenizer,
    compute_metrics=compute_metrics,
)

trainer.train()
```

```
trainer.evaluate()

from transformers import pipeline
tokenizer = AutoTokenizer.from_pretrained("Huffon/klue-roberta-base-nli")
tokenizer.sep_token

# initialize the classifier
classifier = pipeline(
    "text-classification",
    model="Huffon/klue-roberta-base-nli",
    return_all_scores=True,
)

classifier(f"호스트분은 영어밖에 못하십니다. {tokenizer.sep_token} 호스트분도 엄청 친절하시고 영어도 잘하십니다.")
```

6) Evaluate the model, Make a Prediction

```
[[{'label': 'ENTAILMENT', 'score':
0.00026854651514440775},
{'label': 'NEUTRAL', 'score':
0.0006742397672496736},
{'label': 'CONTRADICTION', 'score':
0.999057233335876}]]
```

Highest score = model thinks that label is correct

3. Research Method

3.2. Procedure

➤ Output

premise	new_hypothesis	label
시내랑은 걸어서 걸어서 시내에 갈 방법 있습니다.	시내랑은 걸어서 걸어서 시내에 갈 방법 있습니다.	0
시내랑은 걸어서 걸어서 시내에 갈 숫자 있습니다.	시내랑은 걸어서 걸어서 시내에 갈 숫자 있습니다.	1
또한 스크린 캡처신분증 위조에 사용하기 어렵도록 하는 기량이 있다.	또한 스크린 캡처신분증 위조에 사용하기 어렵도록 하는 기량이 있다.	0
또한 스크린 캡처신분증 위조에 사용하기 어렵도록 하는 쓰임이 있다.	또한 스크린 캡처신분증 위조에 사용하기 어렵도록 하는 쓰임이 있다.	0
송영호 씨와 같은 송영호 씨와 같은 정기 헌혈자의 방법은 점차 감소하고 있다.	송영호 씨와 같은 송영호 씨와 같은 정기 헌혈자의 방법은 점차 감소하고 있다.	1
송영호 씨와 같은 송영호 씨와 같은 정기 헌혈자의 숫자는 점차 감소하고 있다.	송영호 씨와 같은 송영호 씨와 같은 정기 헌혈자의 숫자는 점차 감소하고 있다.	1
신청자 중 사업여좌난 감정 100명이 신청했다.	신청자 중 사업여좌난 감정 100명이 신청했다.	2
신청자 중 사업여좌난 100명이 신청했다.	신청자 중 사업여좌난 100명이 신청했다.	2
한글 창제 과정북한글의 역사에 대해 생생하게 접할 방법 있게 해주는 유물들도 있	한글 창제 과정북한글의 역사에 대해 생생하게 접할 방법 있게 해주는 유물들도 있	0
한글 창제 과정북한글의 역사에 대해 생생하게 접할 숫자 있게 해주는 유물들도 있	한글 창제 과정북한글의 역사에 대해 생생하게 접할 숫자 있게 해주는 유물들도 있	1
영화도 끝날 때 7영화도 끝까지 봐야 그 진가를 알 방법 있다.	영화도 끝날 때 7영화도 끝까지 봐야 그 진가를 알 방법 있다.	0
영화도 끝날 때 7영화도 끝까지 봐야 그 진가를 알 숫자 있다.	영화도 끝날 때 7영화도 끝까지 봐야 그 진가를 알 숫자 있다.	1
경주의 경제구조 경주에서 오래된 마켓은 시내 한복판에 자리하고 있다.	경주의 경제구조 경주에서 오래된 마켓은 시내 한복판에 자리하고 있다.	0
경주의 경제구조 경주에서 오래된 시의 책임자는 시내 한복판에 자리하고 있다.	경주의 경제구조 경주에서 오래된 시의 책임자는 시내 한복판에 자리하고 있다.	2
경주의 경제구조 경주에서 오래된 시장은 시내 한복판에 공간하고 있다.	경주의 경제구조 경주에서 오래된 시장은 시내 한복판에 공간하고 있다.	0
경주의 경제구조 경주에서 오래된 시장은 시내 한복판에 못자리하고 있다.	경주의 경제구조 경주에서 오래된 시장은 시내 한복판에 못자리하고 있다.	1
이번 행사는 관공이변 식은 관광객들에게 공연관광을 알리기 위해서 마련됐다.	이번 행사는 관공이변 식은 관광객들에게 공연관광을 알리기 위해서 마련됐다.	0
이번 행사는 관공이변 발들은 관광객들에게 공연관광을 알리기 위해서 마련됐다.	이번 행사는 관공이변 발들은 관광객들에게 공연관광을 알리기 위해서 마련됐다.	0

premise	new_hypothesis	label	output
시내랑은 걸어서 걸어서 시내에 갈 방법 있습니다.	시내랑은 걸어서 걸어서 시내에 갈 방법 있습니다.	0	0
시내랑은 걸어서 걸어서 시내에 갈 숫자 있습니다.	시내랑은 걸어서 걸어서 시내에 갈 숫자 있습니다.	1	0
또한 스크린 캡처신분증 위조에 사용하기 어렵도록 하는 기량이 있다.	또한 스크린 캡처신분증 위조에 사용하기 어렵도록 하는 기량이 있다.	0	0
또한 스크린 캡처신분증 위조에 사용하기 어렵도록 하는 쓰임이 있다.	또한 스크린 캡처신분증 위조에 사용하기 어렵도록 하는 쓰임이 있다.	0	0
송영호 씨와 같은 송영호 씨와 같은 정기 헌혈자의 방법은 점차 감소하고 있다.	송영호 씨와 같은 송영호 씨와 같은 정기 헌혈자의 방법은 점차 감소하고 있다.	1	1
송영호 씨와 같은 송영호 씨와 같은 정기 헌혈자의 숫자는 점차 감소하고 있다.	송영호 씨와 같은 송영호 씨와 같은 정기 헌혈자의 숫자는 점차 감소하고 있다.	1	1
신청자 중 사업여좌난 감정 100명이 신청했다.	신청자 중 사업여좌난 감정 100명이 신청했다.	2	1
신청자 중 사업여좌난 100명이 신청했다.	신청자 중 사업여좌난 100명이 신청했다.	2	1
한글 창제 과정북한글의 역사에 대해 생생하게 접할 방법 있게 해주는 유물들도 있	한글 창제 과정북한글의 역사에 대해 생생하게 접할 방법 있게 해주는 유물들도 있	0	0
한글 창제 과정북한글의 역사에 대해 생생하게 접할 숫자 있게 해주는 유물들도 있	한글 창제 과정북한글의 역사에 대해 생생하게 접할 숫자 있게 해주는 유물들도 있	1	0
영화도 끝날 때 7영화도 끝까지 봐야 그 진가를 알 방법 있다.	영화도 끝날 때 7영화도 끝까지 봐야 그 진가를 알 방법 있다.	0	0
영화도 끝날 때 7영화도 끝까지 봐야 그 진가를 알 숫자 있다.	영화도 끝날 때 7영화도 끝까지 봐야 그 진가를 알 숫자 있다.	1	0
경주의 경제구조 경주에서 오래된 마켓은 시내 한복판에 자리하고 있다.	경주의 경제구조 경주에서 오래된 마켓은 시내 한복판에 자리하고 있다.	0	0
경주의 경제구조 경주에서 오래된 시의 책임자는 시내 한복판에 자리하고 있다.	경주의 경제구조 경주에서 오래된 시의 책임자는 시내 한복판에 자리하고 있다.	2	1
경주의 경제구조 경주에서 오래된 시장은 시내 한복판에 공간하고 있다.	경주의 경제구조 경주에서 오래된 시장은 시내 한복판에 공간하고 있다.	0	0
경주의 경제구조 경주에서 오래된 시장은 시내 한복판에 못자리하고 있다.	경주의 경제구조 경주에서 오래된 시장은 시내 한복판에 못자리하고 있다.	1	1
이번 행사는 관공이변 식은 관광객들에게 공연관광을 알리기 위해서 마련됐다.	이번 행사는 관공이변 식은 관광객들에게 공연관광을 알리기 위해서 마련됐다.	0	0
이번 행사는 관공이변 발들은 관광객들에게 공연관광을 알리기 위해서 마련됐다.	이번 행사는 관공이변 발들은 관광객들에게 공연관광을 알리기 위해서 마련됐다.	0	0

➤ Output column is the result of training.

➤ The accuracy is obtained by comparing the output column with the label column, which is the correct answer column.

4. Analysis Results

4.1. Accuracy

- Compared to inputting 400 sentences of the original KLUE NLI dataset, as a result of inputting 969 sentences of the expanded dataset, it was confirmed that the accuracy of the model dropped from 0.992481203007518 to 0.7458677685950413.

```
import pandas as pd

# 데이터 파일 로드
df = pd.read_excel("data.xlsx", sheet_name="원본")

# 정답과 예측값 추출
label = df["label"]
output = df["output"]

# 정확도 계산
accuracy = (label == output).mean()

# 결과 출력
print("정확도:", accuracy)

정확도: 0.9924812030075187
```

```
[ ] import pandas as pd

# 데이터 파일 로드
df = pd.read_excel("data.xlsx", sheet_name="확장")

# 정답과 예측값 추출
label = df["label"]
output = df["output"]

# 정확도 계산
accuracy = (label == output).mean()

# 결과 출력
print("정확도:", accuracy)

정확도: 0.7458677685950413
```

- This can be said to have succeeded in fooling the model.

4. Analysis Results

4.2 Confusion Matrix

Original KLUE NLI dataset

label \ output	0	1	2
0	139	0	0
1	1	125	0
2	0	2	132

Extended dataset

label \ output	0	1	2
0	158	41	8
1	49	317	58
2	15	75	247

- From the results of running the [extended dataset](#), the model recognized incorrectly in the following order:
 - contradiction is recognized as neutral
 - neutral is recognized as contradiction
 - neutral is recognized as entailment
 - entailment is recognized as neutral
 - contradiction is recognized as entailment
 - entailment is recognized as contradiction

5. Conclusion

5.1. Summary

- The purpose of adversarial examples is to trick the model by deliberately adding noise to the data.
- In order to create adversarial cases that can be used to train a model for analyzing the implication relationship between texts, hypothesis sentences from the original dataset were taken and noun words in the sentences are replaced with synonyms of their homonyms.
- By running the word replacement algorithm, a newly extended dataset (consisting of premise-hypothesis-implication relationship) is obtained.
- As a result of training the expanded dataset with the Roberta model, the accuracy was reduced.
- In the case of errors, the most common case was to misidentify the contradiction as neutral.
- As a result, we succeeded in fooling the model.
- Therefore, it can be expected that a homonyms-based word replacement algorithm will help generate adversarial examples.

6. References

- Kim, Jae Kyun. Copus Augmentation Using Synonym Replacement, Insertion, Deletion: A Focus on Korean Named Entity Tagged Corpus. MA thesis. Korea Maritime & Ocean University, 2021.
- Lee, Y. 2022. Adversarial Implications Analysis Explanatory Material. Daejeon: Chungnam National University.
- Lee, Jaehun. “생성적 적대 신경망 소개.” Pasetul’s Devlog. 2019. Web. 27 May 2023.
- “Korean Vocabulary Learning List.” National Institute of Korean Language, 2003. Web. 27 May 2023.
- Park, sungjoon and Others. 2021. KLUE: Korean Language Understanding Evaluation. arXiv preprint. arXiv: 2105.09680
- Han, Jiyoung, et al. “KLUE-NLI-Natural Language Inference.” Klue-benchmark. 2021. Web. 27 May 2023.
- Lee, Y. 2023. Applied Linguistics Lecture Note. Daejeon: Chungnam National University.

A Corpus-based Multi-lingual Comparison of AI-based Machine Translation: English, Korean, and Chinese

Cuilin Liu, Homin Park, Se-Eun Jhang

(Korea Maritime & Ocean University)

June 3, 2023

The KACL-KASELL Summer Joint Conference, Korea University, ROK

Contents

1. Introduction

- Research Background
- Implications of the Ideas
- Research Questions

2. Literature Review

3. Methods

- How to Calculate TTR
- Research Steps including Evaluation Programs of Machine Translation
- Tools

4. Results and Discussion

5. Conclusion

References

1. Introduction

- The present study aims to examine a quantitative method (TTR) in identifying the discrepancy of the productions of Machine Translation Systems (MTS).
- As a text becomes longer, **the type-token ratio becomes lower**. This is because the number of word-tokens continues to rise while the number of word-types rises more and more slowly.
- One reason is that the speaker-writers vocabulary is **finite**.
- The speaker-writers must continually choose between **‘old’ words** or **‘new’ words** consciously or influenced by grammatically rules (Tuldava, 1998).

1. Introduction (“New words and new topics”)

- The speaker-writers need to consider two opposing matters:
 - **new words** are needed to develop and **broaden the topic**
 - **old words** are needed to make the **text cohesive** (Stubbs, 2001)

So, the speaker-writers make choices from the available vocabulary, alternating between **repeating old words and introducing new words**, which not only influence the whole texts but also **the smaller sections of texts**.

These ideas imply that it is plausible to **predict new topics in essays or new episodes in stories by detecting alternations of new and old vocabularies**.

1. Introduction (“Implications of the ideas”)

The above ideas about TTR have implications for topics such as measures of reading difficulty and the amount of **given information** in discourse, **identifying the discourse boundaries** and estimating how adequately a corpus represents a language.

Youmans (1991) has proposed a computational method of text analysis based on these ideas. By sampling the type-token ratio within **a moving span** of running word-tokens, the author’s program will **identify points in the text where new words** are used for the first time.

Stubbs (2001) has adjusted this method to detect the **significant boundaries** of the short story *Eveline* written by James Joyce, by setting the span **at 151**.

1. Introduction (“Present study & research questions”)

- The present study will apply this technique to identify the **significant boundaries** both in English original text and the **Chinese and Korean translated text** generated by different AI-based MT systems. The assumption is that, since the idea about TTR can be applied in English texts, it should be plausible to **predict new episodes in its corresponding Chinese and Korean translated versions**. It will then shed some light on the **discrepancy** of the translated texts generated by different MTS.
- Research Questions
 - Q1. Whether or not the translated Chinese and Korean TTR curves correlate with the three discourse boundaries claimed by literary critics?
 - Q2. Do their TTR graphs show any discrepancies in the three AI-based machine translation systems such as Google Translate, DeepL, and GPT 3.5? How do we evaluate which is better or worse?

2. Literature Review (“Two main directions in using TTR”)

- Two main streams in using TTR:
- **Overall TTR (or STTR)** is commonly applied in English overall text analysis based on lexical diversity:

e.g., Measuring reading difficulty

Authorship attribution

Corpus representativeness

Lexical diversity of human translations and outputs of MTS

‘**Separate TTR**’ is used in discourse analysis based on segments of text:

- e.g., 1. Information flow – given/accessible information, new information
2. Discourse constituents

2. Literature review (“Studies related to machine translations and lexical diversity (using overall TTR)”)

- Toral (2019) studied lexical density and diversity and found that post-edited machine translations, compared with human translations
- Vanmassenhove, Shterionov, and Way (2019) studied the output of 12 different machine translation systems with original and back-translated data.
- Vanmassenhove, Shterionov, and Gwilliam (2021) examined the lexical diversity of MTS by comparing training data (human translations) and the output of machine translation systems trained on the same dataset using different architectures (phrase-based statistical system, neural long short-term memory (LSTM) network, and neural transformers).

2. Literature Review (“Studies calculating TTR separately”)

Youmans (1991) extended Chafe’s (1987) categories of **three informativity** – **given, accessible, new information** by analogy to **vocabulary management** in text.

According to Youmans, **repeated words** might be classified as ‘given information’, **new function words** as ‘accessible information’, and **new content words** as ‘new information’.

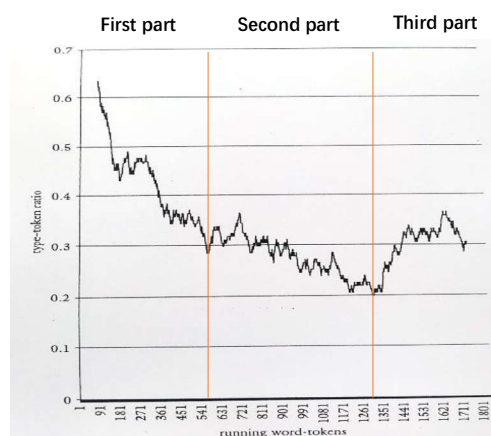
To visualize the rich rhythmic variations, i.e, to detect the ebb and flow of new vocabulary and show their **boundaries** clearly, he proposed a new method called **vocabulary-management-profiles** (VMPs).

- There is a rhythmic alternation between **new and repeated** vocabulary in the typical well-crafted story (Youmans 1991, span=35).
- No obvious peaks and valleys (span=101). **They can’t detect patterns shorter than the span. In other words, longer spans can generate smoother VMPs.**

2. Literature Review (“Stubbs 2001”)

- Stubbs (2001) applied the method proposed by Youmans (1991) to examine the correspondence of **the text boundaries** in the **TTR graph** with the **literary critics** made by literature experts.

Literary critical interpretation vs. TTR (*Eveline* by James Joyce)



Eveline, span 151

	First Part	Second Part	Third Part
Hart (1969)	Eveline thinks about the past and possible future	Brief interlude in which she reasserts her decision to choose life	Third part ends in her psychological failure
Stubbs (2001)	<ul style="list-style-type: none"> • Up-slope: High TTR means that a group of new vocabulary is introduced to make a new turn in the story. • “These three parts correspond astonishingly closely to the troughs and peaks identified by the program” • “The literary critic’s impressions are quite correct, but they can now be given an objective textual basis.” 		

2. Literature Review (“Significant boundary and the present study”)

- According to Stubbs (2001), if a **prominent up-slope** occurs late in the text, it is likely to show a significant boundary. By concentrating on **the lowest point and the following single most prominent rise in the curve**, he found the correspondence of the lowest point to the late sentence which shows most identical words to the opening sentence.
- Based on our knowledge, despite some linguists studies MT using the TTR technique, there seems little study using TTR by calculating it separately, and even less testing this method **in translated texts either by human or MTS**.
- The present study will extend the use of VMPs by examining the validation of the VMPs in revealing the discrepancy of **Korean and Chinese translations** both by human and MTS.

3. Methods (“Basic definitions related to TTR”)

The method applied in the present study is to calculate the type-token ratio **separately for different segments** of text, by sampling the type-token ratio within **a moving span** of running word-tokens.

I. Counting words

- **English words:** the present study identifies the words as the *graphic tokens*-word divided by spaces; only hyphens and apostrophe are included. (Francis & Kučera, 1982)
- **Chinese words & Korean words:** Both the Chinese translated texts and Korean translated texts are tagged with part of speech and then tokenized as words, which are calculated as word-tokens (e.g., Chao, 1968; Norman, 1988).

II. Counting TTR

Moving span: The moving span proceeds token by token through the text, e.g., from word 1 to 151, from 2 to 152, and so on. For each span, the program calculate the ratio of new words and old words and the ratio is stored (Stubbs, 2001).

3. Methods (“Pilot study about the span”)

The span is **an intuitive decision** with the aim to obtain interesting results which can reveal significant boundaries within text. If the span is too short, no detail and no clear patterns emerge. If it is too long, it will be too smooth (Stubbs, 2023, personal communication).

It seems that the decision of the span is related to **the length of the novel**.

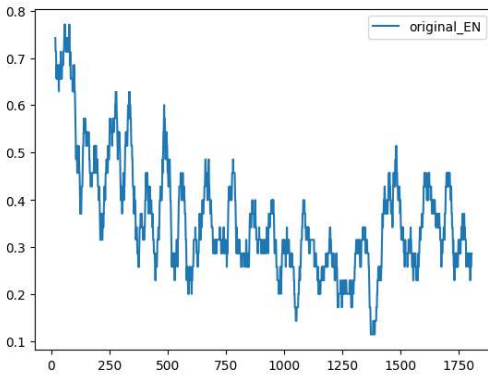


Fig.1 Original English *Eveline*, span=35

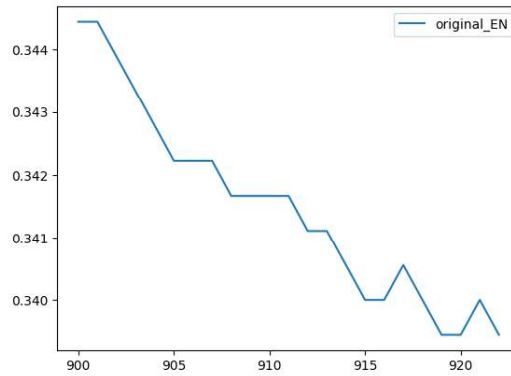


Fig.2 Original English *Eveline*, span=1800

3. Methods (“Three metrics for evaluation of the MT outputs”)

- In the present study, **comprehensive metrics** (BLEU scores, METEOR, ROUGE) have been adopted for evaluating the quality of MT productions which were then compared with the results obtained from TTR.
- As for the benchmark of the three metrics, they all require human-generated translations as a standard basis for comparison. So both Chinese and Korean translations of *Eveline* were selected from professional translators.
 - Chinese version is translated by Wang, Fengzhen
 - Korean version is translated by Han, IL-Dong
- The scores produced by these metrics are typically **normalized to a scale between 0 and 1** for easier interpretation. A score of **1 indicates a perfect match** or similarity, while a **score of 0 indicates no overlap or similarity**.

3. Methods (“Major steps”)

- Getting TTR results for the original English *Eveline*, as well as for human and machine-translated *Eveline* (Chinese & Korean)
- Getting the scores of the three MTS using metrics of BLEU, METEOR and ROUGE.
- Comparing the average scores of the three metrics with the TTR graphs for each MT model, so as to assess the validation of the TTR technique.
- Take a closer observation of the TTR graphs to make comparisons between original English version and all translated versions.

3. Methods (“Tools: Python”)

ANACONDA: <https://www.anaconda.com/>

- Jupyter: It can be used through the website or in the Vscode.
- Python+Vscode: www.python.org + <https://code.visualstudio.com/>

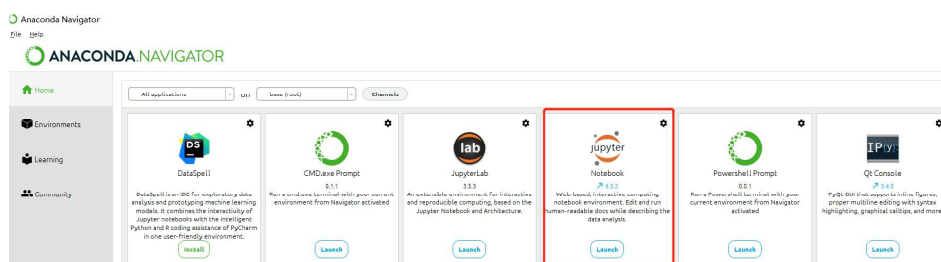


Fig.3 Website to get access to Jupyter

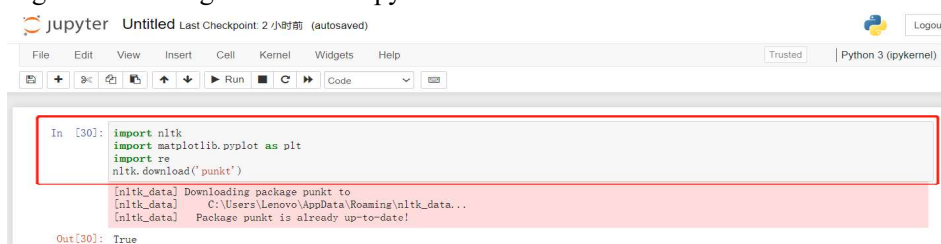


Fig.4 Jupyter input cell

Note:

For Chinese and Korean POS tagging, TagAnt 2.0.5 was used for Chinese translated texts and [KiwiGui v0.15.0](https://github.com/bab2min/Kiwi#%EC%9D%B8%EC%9A%A9%E D%95%98%EA%B8%B0) for Korean translated texts.

POS tagging for Korean:

<https://github.com/bab2min/Kiwi#%EC%9D%B8%EC%9A%A9%E D%95%98%EA%B8%B0>

POS tagging for Chinese:

<https://www.laurenceanthony.net/software/tagant/>

4. Results and Discussion (“TTR separately counted for Chinese and Korean human translations”)

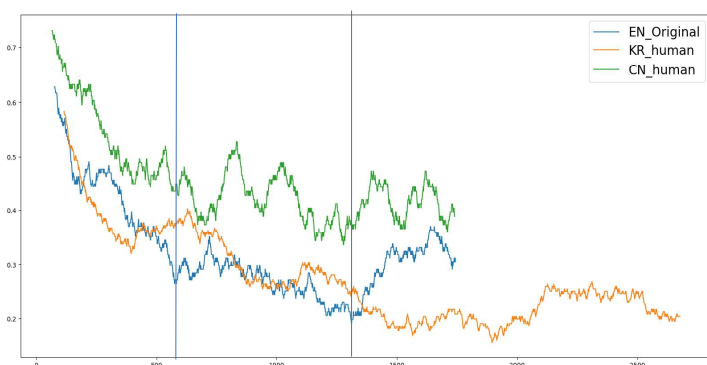


Fig. 5 English *Eveline*, Korean Human Translation, Chinese Human Translation

EN: span=151, total number of words: 1,820

KR: span=231, total number of words: 2,794

CH: span=131, total number of words: 1,807

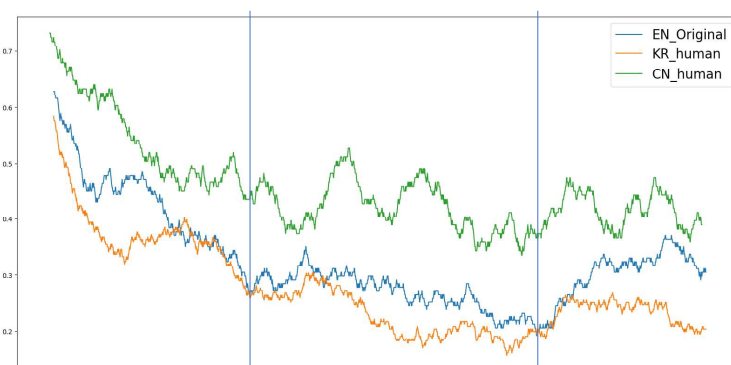


Fig. 6 English *Eveline*, Korean Human Translation, Chinese Human Translation (Normalized comparison)

Fig. 5 and Fig.6 show that both Korean and Chinese human translations have relatively clear two main lowest points, which is important to indicate the two significant boundaries. Since the total number of each text is different from each other, the two lowest points don't coincide with that of English version. In the normalized graph, it can be seen that the two lowest points in two lines are very near the those of the original English version.

4. Results and Discussion (“Three MTSs in Korean”)

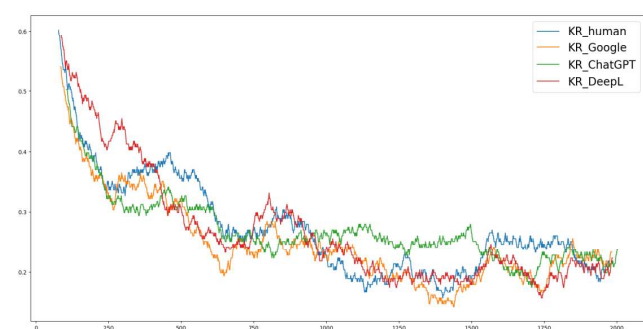


Fig. 7 TTR chart of Google Translate

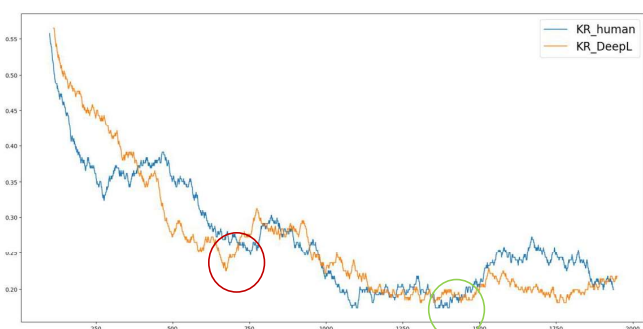


Fig. 8 TTR chart of DeepL

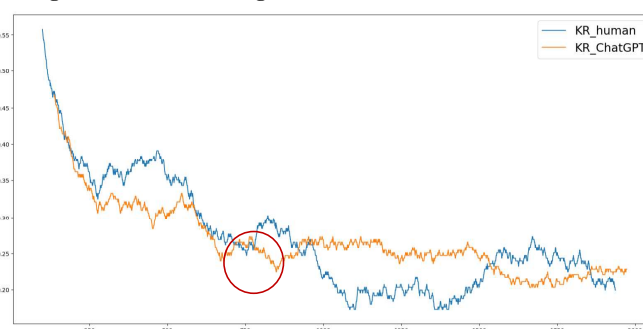


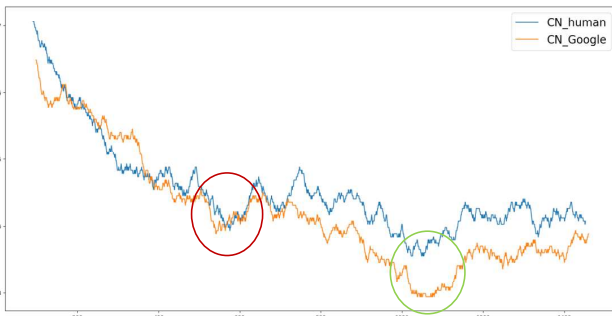
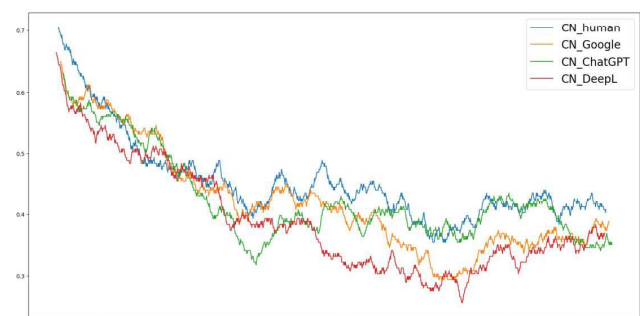
Fig. 9 TTR chart of GPT Playground

Discrepancy:

Resemblance between the line of human translations and MTSs:

DeepL >
Google >
GPT

4. Results and Discussion (“Three MTSs in Chinese”)



Discrepancy:
Resemblance between the line of human translations and MTs:
GPT > Google > DeepL

Fig. 10 TTR chart of Google Translate

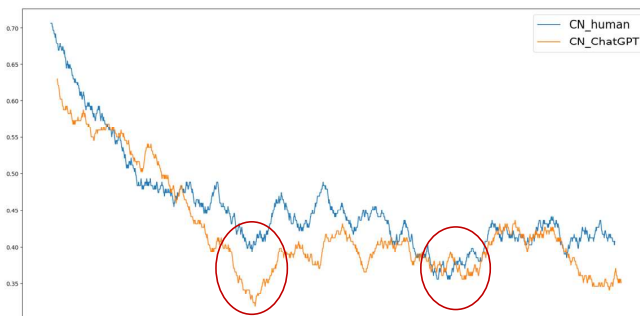


Fig. 11 TTR chart of ChatGPT

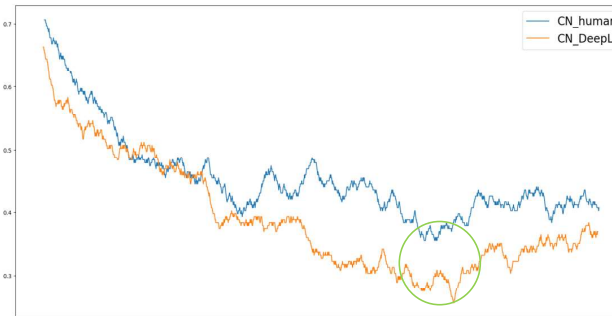


Fig. 12 TTR chart of DeepL

4. Results and Discussion

Table 1. Evaluation of different metrics, Korean outputs of MTS

Korean	Google	DeepL	GPT 3.5
BLEU	0.2167 (3 rd)	0.2402 (1 st)	0.2213 (2 nd)
METEOR	0.3395 (2 nd)	0.3549 (1 st)	0.3176 (3 rd)
ROUGE	0.4347 (2 nd)	0.4438 (1 st)	0.3965 (3 rd)
Ranking	2nd	1st	3rd

Korean Ranking:
DeepL > Google > GPT

Chinese Ranking:
GPT > Google > DeepL

It seems that different MT models reveal different quality in particular languages.

Table 2. Evaluation of different metrics, Chinese outputs of MTS

Chinese	Google	DeepL	GPT 3.5
BLEU	0.1873 (1 st)	0.1790 (3 rd)	0.1795 (2 nd)
METEOR	0.3523 (2 nd)	0.3444 (3 rd)	0.3606 (1 st)
ROUGE	0.4971 (2 nd)	0.4767 (3 rd)	0.5156 (1 st)
Ranking	2nd	3rd	1st

In other words, it seems that AI-based MTSs are sensitive to different languages

5. Conclusion

Q1. Whether or not the translated Chinese and Korean TTR curves correlate with the three discourse boundaries claimed by literary critics?

- The TTR calculated separately with a moving span seems plausible to detect the significant boundaries in Chinese and Korean translations. **It seems that the separately-counted TTR is sensitive to show the significant parts of a short story.**

Q2. Do their TTR graphs show any discrepancies in the three AI-based machine translation systems such as Google Translate, DeepL, and GPT 3.5? How do we evaluate which is better or worse?

- For both Korean and Chinese, the TTR graphs show discrepancies between human translations and the three MTs. For Korean, DeepL's line appears to match human translation the best, as is closer to human for most of the line. Compared to human translation, in the Korean translation TTR line plots, Google looks the second most similar and GPT looks the worst. However, in the Chinese translation TTR line plots, the reverse is true, with ChatGPT looking the best similar, Google looking the second and DeepL looking the worst.
- **This result about Korean and Chinese is the same as the ranking of scores we obtained through three evaluation metrics because the best ranking of the machine translation evaluation for Korean was DeepL, Google, and ChatGPT, while for Chinese it was ChatGPT, Google, and DeepL. It seems that TTR has the potential to reveal the quality differences of MTs but it is also sensitive to particular languages.**

References

- Chafe, W. L. (1987). Cognitive constraints on information flow. In Russell S. Tomlin (Ed.), *Coherence and Grounding in Discourse* (pp. 21–51). Philadelphia: John Benjamins.
- Chao, Y. R. (1968). *A Grammar of Spoken Chinese*. Berkeley and Los Angeles: University of California Press.
- Francis, W. N., and Kučera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin.
- Norman, J. (1988). *Chinese*. Cambridge: Cambridge University Press.
- Stubbs, M. (2001). *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell
- Toral, A. (2019). Post-edits: an exacerbated translationese. In M. Forcada, A. Way, B. Haddow, and R. Sennrich (Eds.), *Proceedings of Machine Translation Summit XVII: Research Track* (pp. 273–281). Dublin: European Association for Machine Translation.
- Tuldava, J. (1998). Probleme und Methoden der quantitativ-systemischen Lexikologie. [Translated from Russian original (1987).] Trier: Wissenschaftlicher Verlag.
- Vanmassenhove, E., Shterionov, D., and Gwilliam, M. (2021). Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translation. In P. Merlo, J. Tiedemann, and R. Tsarfaty (Eds.) *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 2203–2213).
- Vanmassenhove, E., Shterionov, D., and Way, A. (2019). Lost in translation: Loss and decay of linguistic richness in machine translation. In M. Forcada, A. Way, B. Haddow, and R. Sennrich (Eds.), *Proceedings of Machine Translation Summit XVII: Research Track* (pp. 222–232).
- Youmans, G. (1991). A new tool for discourse analysis: the vocabulary management profile. *Language*, 67 (4): 763-89.

A corpus-based analysis of lexical bundles in maritime legal genres

Guandong Zhang
(Korea Maritime & Ocean University)

June 3, 2023

The KACL-KASELL Summer Joint Conference, Korea University, ROK

Abstract



- As an English for specific purposes, maritime English has its own specialized vocabulary, grammar, and communication patterns. However, there is little research focusing on the frequently used lexical bundles in different maritime legal genres. This paper attempts to investigate frequently used lexical bundles in four maritime legal genres, namely case law, legal documents, legislation and academic legal articles, based on a self-built Maritime English Law Corpus. In this study, the text dispersion-based keyword analysis proposed by Egbert and Biber (2019) is used to extract domain-specific lexical bundles as keywords serve as an efficient indicator of such lexical bundles according to Qi (2019). All 2- to 5-word lexical bundles uniquely used in four maritime legal genres are extracted using WordSmith Tools 8.0 (Scott, 2020) and analyzed in terms of syntactic and functional characteristics. Furthermore, a comparative analysis is carried out using Breeze's (2013) criteria to compare the similarities and differences between General English and Specialized English in legal texts, specifically 4-word lexical bundles in the four maritime legal genres. The results reveal significant differences in the types and roles of lexical bundles in the four maritime legal genres, as well as unique linguistic features specific to maritime legal texts.

Key Words: n-grams, 4-word lexical bundles, text dispersion-based keyword analysis, legal genres, maritime English law corpus

1. Research Questions

(1) What are the most frequent lexical bundles across four legal genres?

→ Comparison of general statistics for 2-5 word bundles in four genres will be analyzed through WordSmith Tools 8.0 (Scott, 2020).

(2) What are the homogeneous features of maritime legal genres that distinguish them from general legal texts?

→ Comparison of 4-word lexical bundles in maritime legal texts and general legal texts will be analyzed.

(3) How do we differentiate one genre from the others? In other words, does each of the four maritime English legal genres have its own unique characteristics?

→ Comparison of 4-word lexical bundles lists will be further analyzed to explore characteristics of bundles of each genre.

2. Literature Review

- Breeze (2013) investigated 4-word lexical bundles in four legal corpora: academic law, case law, legislation, and documents.
 - Two million words are composed of each subcorpus containing about 500,000 words.
 - Focusing on major differences and the roles between syntactic and discourse patterns of these bundles in the four legal genres.

2. Literature Review

- Qi (2018) investigates domain-specific multiword units (MWUs) in marine accident investigation reports (MAIR) to characterize their syntactic, semantic, and functional features.
- Majority of domain-specific MWUs consist of two-word sequences.
- Only a small percentage of MWUs occur commonly, while most have low frequencies.
- MAIR genre employs a wide variety of domain-specific MWUs rather than relying on a few common expressions.

3. Data and Methodology

3.1 Data

- **Study corpus:**
 - This current study corpus consists of about 1 million words from maritime law in English, divided into four sub-corpus of approximately 250,000 words for each genre.
 - Case law (admiralty and maritime cases, admiralty court opinion, law reports)
 - Document (charter contracts)
 - Legislation (IMO Conventions)
 - Academic (maritime law journal papers)
- **Reference corpus:**
 - BNC Baby containing four one-million-word genre-based subsets, extracted from the British National Corpus (Academic /Fiction /Newspaper /Conversation)

General statistics of the study corpus

Corpus	Genres	Tokens	Types	Type/Token ratio (TTR)	Standardized TTR	STTR std. dev.
Study Corpus (Maritime English Law Corpus)	Case Law	252,311	10,990	4.57	34.02	64.71
	Document	253,728	4,872	1.95	29.36	69.63
	Legislation	250,018	6,494	2.70	28.38	70.88
	Academic	254,255	14,430	5.85	38.86	59.76
	Total	1,010,312	21,568	2.21	32.77	66.33

7

3.2 Methodology

- WordSmith Tools was used to identify all 2- to 5- lexical bundles in four genres.
 - Setting min frequency of 5/min text of 1
- Text Dispersion-based Keyword Analysis :
 - Words that occur in statistically more texts in a target corpus when compared with a reference corpus (Egbert and Biber, 2019)
 - Setting p-value of 0.1/minimal frequency 3 times
- Keyword-bundles (Qi, 2019)
 - incorporate ‘meaning’ into the statistics-based approach because keywords serve as an efficient indicator of the domain-specific n-grams.

8

4. Results and Discussion

- Number of 2-5 lexical bundles in maritime legal texts

All lexical bundles in four genres

NO	genres	clusters-2	clusters-3	clusters-4	clusters-5	total
1	Case law	5,575	2,456	858	397	9,286
2	Document	6,727	6,879	6,155	5,337	25,098
3	Legislation	6,312	4,712	3,154	2,190	16,368
4	Academic	4,992	1,481	353	135	6,961
Total		23,606	15,528	10,520	8,059	57,713

4. Results and Discussion

- Number of keywords in maritime legal texts

Lemmatization is a sensible approach for productive purposes
 (Nation & Kobeleva, 2016)

Keywords in 4 genres

	Case law	Document	Legislation	Academic	total
Keywords	127	285	224	89	725

• Top 10 and bottom 5 keywords in four genres

N	Case law	BIC	N	Document	BIC	N	Legislation	BIC	N	Academic	BIC
1	court	101.05	1	clause	131.68	1	ship	100.19	1	cannot	91.55
2	cannot	92.95	2	thereof	119.06	2	paragraph	89.72	2	maritime	85.70
3	cir	86.48	3	vessel	116.47	3	annex	88.80	3	eu	42.67
4	pursuant	77.55	4	charterer	114.14	4	accordance	81.49	4	marine	40.60
5	maritime	64.89	5	indemnify	111.08	5	regulation	75.05	5	imo	35.22
6	vessel	61.10	6	accordance	100.86	6	cargo	73.68	6	shipping	34.07
7	judgment	59.37	7	herein	94.91	7	authorized	69.25	7	ship	27.77
8	corp	49.63	8	incurred	85.47	8	comply	65.87	8	coastal	25.27
9	lloyd	46.01	9	bimco	82.33	9	practicable	64.13	9	impact	24.88
10	inc	45.02	10	receipt	80.90	10	provision	60.94	10	stakeholder	24.68
...
123	navigable	3.45	281	malicious	2.76	220	undertake	3.35	85	tonnage	2.73
124	underwriter	3.34	282	transit	2.66	221	nautical	2.93	86	mepc	2.73
125	affirmative	3.21	283	levied	2.63	222	denounced	2.66	87	connectivity	2.73
126	causative	3.21	284	entitle	2.63	223	drill	2.66	88	behavior	2.73
127	paragraph	2.50	285	ballast	2.63	224	longitudinal	2.66	89	emission	2.72

11

4. Results and Discussion

- Number and proportion of keyword-bundles in maritime legal texts
- The chi-square statistic is 1625.6839. The p-value is < 0.00001. The result is significant at $p < .01$.

Keyword-bundles in four genres

genres	clusters-2	clusters-3	clusters-4	clusters-5	total
Case law	608 (10.91%)	415 (16.90%)	208 (24.24%)	101 (25.44%)	1,332 (14.34%)
Document	1,808 (26.88%)	2,509 (36.47%)	2,879 (46.77%)	2,804 (52.54%)	10,000 (39.84%)
Legislation	1,105 (17.51%)	1,215 (25.79%)	1,099 (34.84%)	869 (39.68%)	4,288 (26.20%)
Academic	352 (7.05%)	132 (8.91%)	61 (17.28%)	22 (16.30%)	567 (8.15%)
total	3,873 (23.93%)	4,271 (26.39%)	4,247 (26.24%)	3,796 (23.44%)	16,187 (28.05%)

• Top 5 keyword-bundles in four genres

Case law									
No	clusters-2	Freq.	clusters-3	Freq.	clusters-4	Freq.	clusters-5	Freq.	
1	DISTRICT COURT	303	COURT OF APPEAL	45	DISTRICT COURT ERRED IN	16	JUDGMENT OF THE SUPREME COURT	12	
2	SUPREME COURT	152	COURT HELD THAT	31	DISTRICT COURT DID NOT	16	DECISION OF THE COURT OF	11	
3	COURT OF	83	DISTRICT COURT ERRED	23	BY THE SUPREME COURT	15	COURT OF APPEAL IN THE	10	
4	THIS COURT	61	COURT FOUND THAT	22	COURT OF APPEAL IN	14	DECISION OF THE SUPREME COURT	9	
5	COURT HELD	38	COURT DID NOT	22	DECISION OF THE COURT	11	COURT OF JUSTICE OF THE	8	

Document									
No	clusters-2	Freq.	clusters-3	Freq.	clusters-4	Freq.	clusters-5	Freq.	
1	THIS CLAUSE	380	THIS CLAUSE SHALL	107	PROVISIONS OF THIS CLAUSE	71	PROVISIONS OF THIS CLAUSE THE	42	
2	SUB CLAUSE	181	THIS CLAUSE THE	89	THIS CLAUSE SHALL BE	54	THIS CLAUSE THE ARBITRATION SHALL	28	
3	CLAUSE SHALL	117	ACCORDANCE WITH CLAUSE	86	PURPOSE OF THIS CLAUSE	34	CLAUSE THE ARBITRATION SHALL BE	28	
4	CLAUSE THE	111	UNDER THIS CLAUSE	68	ACCORDANCE WITH SUB CLAUSE	30	THIS CLAUSE SHALL BE FOR	24	
5	WITH CLAUSE	88	CLAUSE SHALL BE	54	THIS CLAUSE THE ARBITRATION	28	CLAUSE SHALL BE FOR THE	24	

13

• Top 5 keyword-bundles in four genres

Legislation									
No	clusters-2	Freq.	clusters-3	Freq.	clusters-4	Freq.	clusters-5	Freq.	
1	SHIP IS	182	THAT THE SHIP	61	WHOSE FLAG THE SHIP	29	WHOSE FLAG THE SHIP IS	23	
2	SHIP SHALL	94	SHIP SHALL BE	49	SAFETY OF THE SHIP	29	SHIP IS ENTITLED TO FLY	21	
3	SHIP AND	77	SHIP WHICH IS	34	LENGTH OF THE SHIP	26	FLAG THE SHIP IS ENTITLED	18	
4	SHIP TO	68	FLAG THE SHIP	29	SIDE OF THE SHIP	24	DATE AND POSITION OF SHIP	18	
5	SHIP OR	68	POSITION OF SHIP	28	FLAG THE SHIP IS	23	STATE WHOSE FLAG THE SHIP	17	

Academic									
No	clusters-2	Freq.	clusters-3	Freq.	clusters-4	Freq.	clusters-5	Freq.	
1	CANNOT BE	64	MARITIME ZONES LEGISLATION	21	LAW FOREIGN MARITIME LIENS	10	FOREIGN LAW FOREIGN MARITIME LIENS	10	
2	THEY CANNOT	5	FOREIGN MARITIME LIENS	20	FOREIGN LAW FOREIGN MARITIME	10	RECOGNITION OF FOREIGN MARITIME LIENS	9	
3	IT CANNOT	5	MARITIME ZONES AND	14	THAT A MARITIME LIEN	9	SPECIAL MARITIME PROCEDURE LAW OF	7	
4	MARITIME ZONES	79	MARITIME SPATIAL PLANNING	14	RECOGNITION OF FOREIGN MARITIME	9	MARITIME PROCEDURE LAW OF CHINA	7	
5	MARITIME LIEN	69	MARITIME LIEN IS	14	LIABILITY FOR MARITIME CLAIMS	8	URBAN AND MARITIME SPATIAL PLANNING	5	

14

4. Results and Discussion

- Breeze's (2013) four major categories of lexical bundles:
 - Content noun phrases and prepositional phrases:
 - refer to specific aspects of the contents of the texts
 - Non-content noun and prepositional phrases:
 - prepositional phrases or prepositional noun phrases of a general nature
 - Adjectival phrases
 - bundles associated with adjectives
 - Verb phrases
 - bundles containing a verb phrase or part of one

15

4. Results and Discussion

- Similarities between maritime legal corpus and general legal corpus

Frequency of non-content prepositional phrase bundles				
4 words-bundles	Case law	Document	Legislation	Academic
AS A MATTER OF	36	0	0	10
ON THE OTHER HAND	12	0	0	31
AS WELL AS THE	11	6	15	40
Frequency of verb bundles in legislation and documents				
BUT NOT LIMITED TO	0	94	8	0

16

4. Results and Discussion

- Xu's (2021) two categories of maritime English words: academic / technical words
 - Academic keyword-bundles in maritime English legal genres:
 - keyword-bundles contain words that occur frequently across a wide range of academic texts but infrequently in other genres
 - Technical keyword-bundles in maritime English legal genres:
 - keyword-bundles contain words that are used and known mainly in a specific subject area, but not others

17

4. Results and Discussion

Classification of keyword-bundles in four genres

	Case law	Document	Legislation	Academic
Academic keyword-bundles	555 (41.67%)	5,596 (55.96%)	2,345 (54.68%)	397 (70.02%)
Technical keyword-bundles	177 (13.28%)	1,441 (14.42%)	643 (15.00%)	59 (10.41%)
others	600 (45.05%)	2,963 (29.63%)	1,300 (30.32%)	111 (19.57%)
total	1,332 (100%)	10,000 (100%)	4,288 (100%)	567 (100%)

18

5. Conclusion

- The proportion of bundles with lengths ranging from 2 to 5 in the total number of bundles exhibits minimal variation. Specifically, the ratios of 2-clusters and 5-clusters are the closest, while the remaining two ratios are the most similar.
- Based on examples provided in the general legal corpus, only a few bundles belonging to the categories of non-content prepositional phrase bundles and verb bundles, and no bundles from the other two categories were identified in the maritime legal corpus.
- According to Xu's classification of academic vs. technical words, **document genre** exhibits the highest occurrence of both academic and technical keyword-bundles, while **academic genre** has the lowest, as unsurprisingly expected in the order of genres in the number of bundles and keyword-bundles.

References

- Breeze, Ruth. 2013. Lexical bundles across four legal genres. *International Journal of Corpus Linguistics*, 18(2), 229-253.
- Egbert, Jesse and Biber, Douglas. 2019. Incorporating text dispersion into keyword analyses. *Corpora*, 14 (1), 77-104.
- Nation, I. S. P., and Kobeleva, P. 2016. Proper nouns. In I. S. P. Nation (Ed.), *Making and using wordlists for language learning and testing* (pp. 55-64). Benjamins.
- Qi, Yilian. 2019. *Syntactic and semantic patterns of domain-specific multiword units in marine accident investigation reports*. Unpublished doctoral dissertation, Korea Maritime and Ocean University.
- Scott, Mike. 2020. *WordSmith tools version 8: Lexical Analysis Software*. Retrieved April 15, 2020 from <https://www.lexically.net/wordsmith/downloads/>
- Xu, Lin. 2021. *Developing vocabulary lists in specialized Maritime English corpora*. Unpublished doctoral dissertation, Korea Maritime and Ocean University.

초등영어 학습자 어휘 수준에 적합한 아동문학 작품 탐색 연구

성민창(경인교육대학교), 김은형(서울대학교),
김기택 (서울대학교)

발표 순서

- 연구의 동기
- 이론적 바탕
- 학습자 수준
- 아동문학작품 분석

EFL 환경

□ 문제의식: 부족한 외국어 입력 시간

<표 3> 우리나라 각 교육단계 별 영어교육 시간

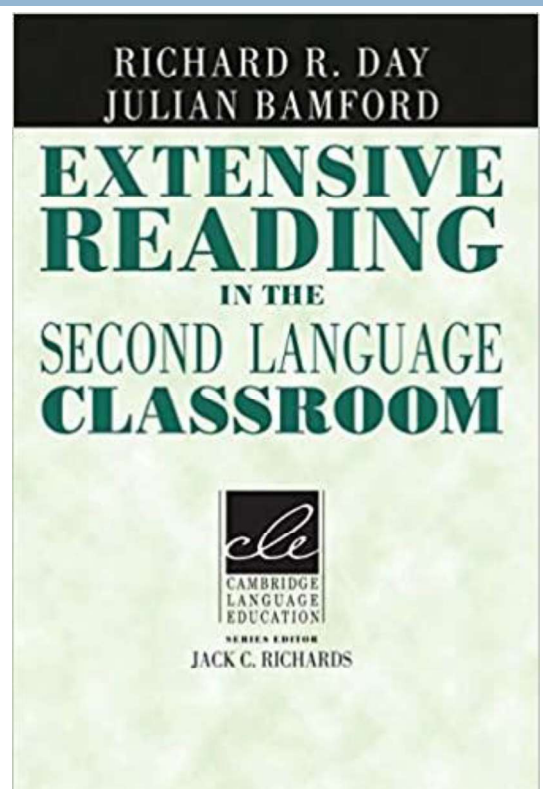
학년단위	주당 시간	년간 시수/ 총 시간
초등학교 3학년	1	34/22.6 시간
초등학교 4학년	1	34/22.6 시간
초등학교 5학년	2	68/45.2 시간
초등학교 6학년	2	68/45.2 시간
중학교 1학년	3	102/76.5 시간
중학교 2학년	3	102/76.5 시간
중학교 3학년	4	136/102 시간
고등학교 1학년	4	136/113.3 시간
고등학교 2학년	4	136/113.3 시간
고등학교 3학년	4	136/113.3 시간
총 영어교육 시간		730.5 시간

이병민, 2003, p. 119

EFL 환경

□ 문제 해결: 입력 양 ↑

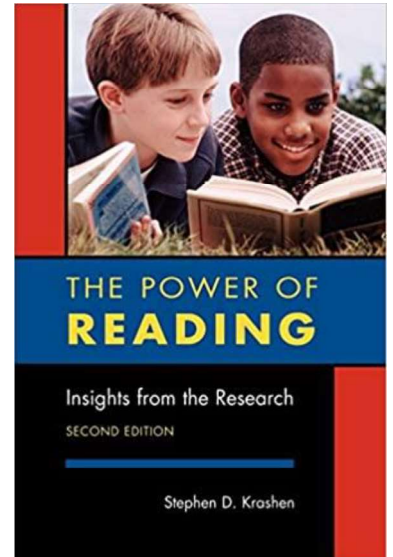
- ▣ 몰입교육, 영어마을...
- ▣ 다독/다청



EFL 환경

□ 결과: '긍정적'

- ▣ 예: 다독의 긍정적 효과 (예: Jeon, 2008)



실천

□ 다독/다청의 실천

- ▣ 누가
- ▣ 언제
- ▣ 어디서
- ▣ 어떻게
- ▣ 왜



실천

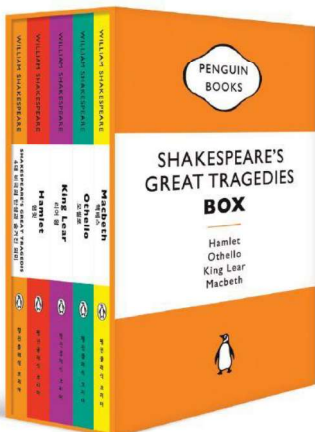
□ 다독/다청의 실천

- 누가
- 언제
- 어디서
- 어떻게
- 왜
- 무엇을...?



무엇을...?

□ 소설, 팝송, 영화...



무엇을...? 정보의 바다

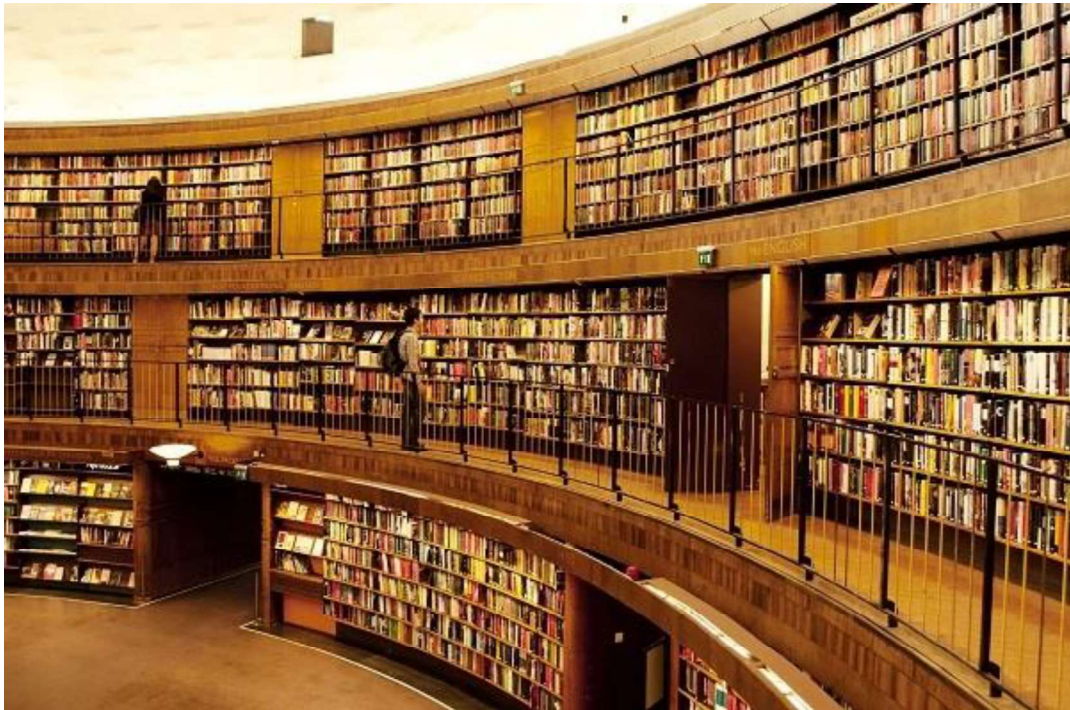
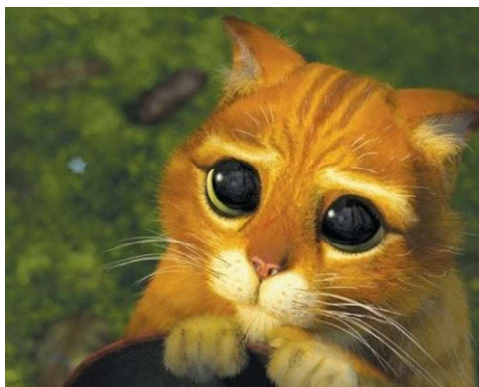


사진 출처: <https://www.readersnews.com/news/articleView.html?idxno=99913>

2023년 현재

“선생님, 무엇을 읽으면 좋을까요?”

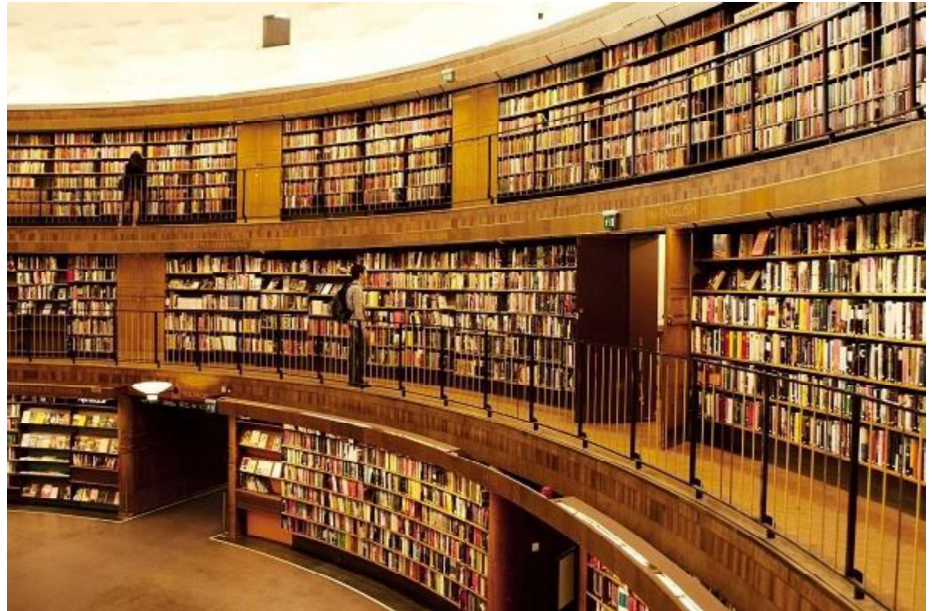
“선생님, 무엇을 들으면 좋을까요?”



2023년 현재

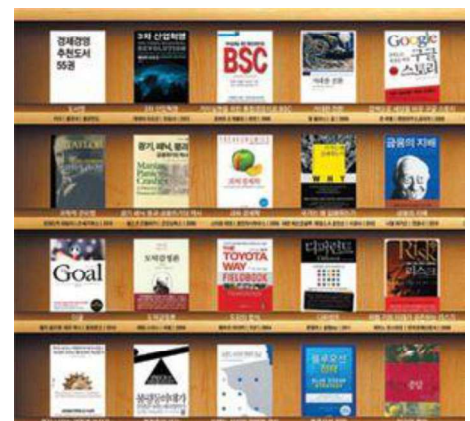
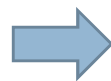
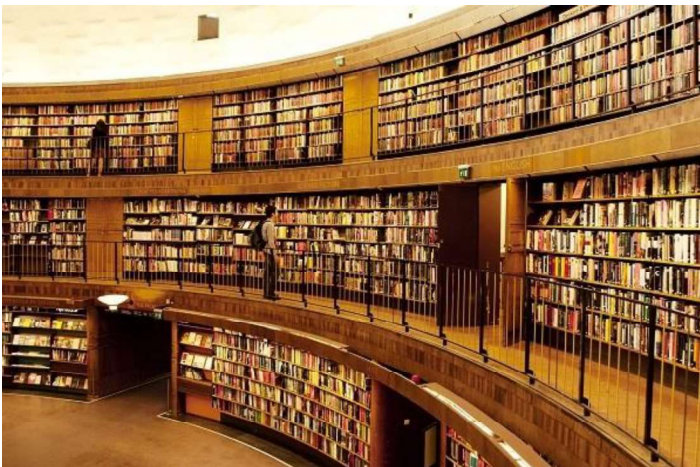
e.g.,
"네 수준에 맞는 것 중에서 흥미있는 것을 읽고
들으렴."

흠.....



목표

- 정보의 바다 속에서, 학습자 수준에 맞는 다독/다청 자료를 추려내 줄 수는 없을까?



발표 순서

- 연구의 동기
- 이론적 바탕
- 학습자 수준
- 아동문학작품 분석

이론적 바탕: 외국어 입력의 중요성

- 언어습득 연구 (Chomsky, 2005; O'Grady, 2012)

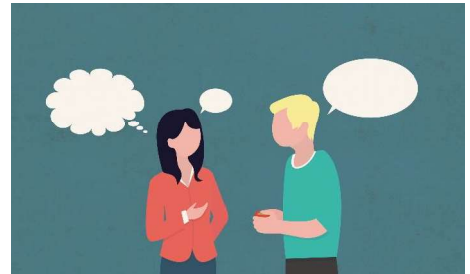
	UG	Emergentism
Experience	○	○
Language-independent principles	○	○
Universal Grammar	○	X

학습자 수준에 적합한 입력

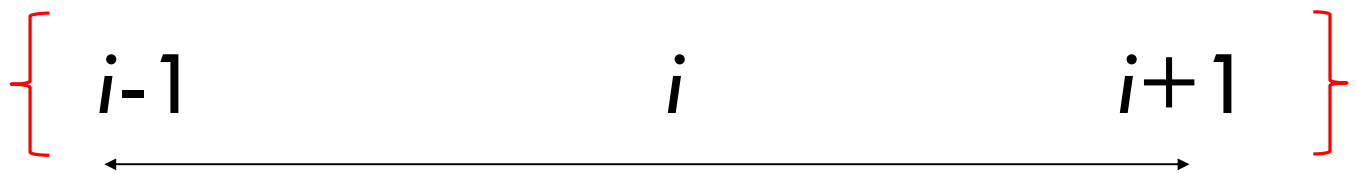
Child-directed speech (Snow & Ferguson 1977)



Foreigner talk (Long 1983)



학습자 수준에 적합한 입력



(Day & Bamford, 1998)

(Krashen, 1982)

학습자 수준에 적합한 입력

- 주어진 텍스트를 스스로 이해할 수 있는데 (unassisted comprehension) 필요한 어휘 규모
 - Token 기준
 - ▣ 98% (e.g., Hu and Nation 2000)
 - ▣ 95% (e.g., Liu and Nation 1985)
 - ▣ 90% (영화; Webb and Rodgers 2009)
- 모르는 단어가 나오더라도 맥락이나 **시각적 정보**를 통해 유추 가능

정리

- 학습자 수준에 적합한 다독/다청 자료란?
 - ▣ 자료의 언어적 수준이 학습자의 현재 수준()보다 약간 높거나 낮음
 - ▣ 자료에 제시된 외국어 입력의 90~98%(token 기준)가 학습자가 이미 알고 있는 어휘로 구성

문제

학습자의 현재 수준()을 어떻게 설정할 것인가?



사진 출처: <https://www.yna.co.kr/view/AKR20181212148500062>

발표 순서

- 연구의 동기
- 이론적 바탕
- 학습자 수준
- 아동문학작품 분석

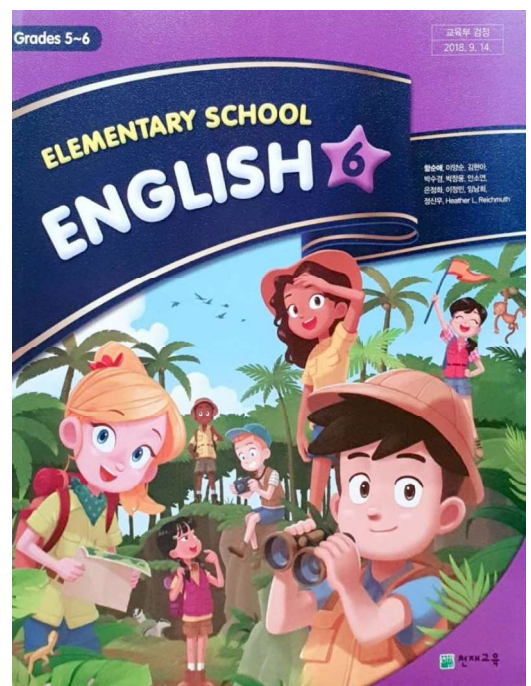
학습자의 현재 수준(i) 설정

- 교실내 모든 학생들의 현재 영어 수준(i)을 파악하는 것은 현실적으로 어려움
- 학교 영어수업을 성실히 이수한 학생의 대략적인 수준($i-1 \sim i+1$)을 가늠하는 것은?

학습자의 현재 수준(i) 설정

교육부 고시 제2015-74호 [별책 14]

영어과 교육과정



학습자의 현재 수준(/) 설정

□ 영어과 교육과정: 어휘

기본 어휘 목록

A

a *

abandon

able

aboard

abort **

about *

above *

abroad

accurate

accuse

achieve

acid

acknowledge

acquaint **

acquire

acquisition **

across *

act *

adverse **

advertize / advertise

advise *

advocate

aesthetic **

affair

affect

affection

affiliate **

afford

raid *

ter *

temoon *

gain *

paint *

7. 기본 어휘 목록에 제시한 어휘는 모두 3,000개이며, 어휘 뒤에 *표를 한 어휘는 800개 로 초 등과정에서 사용하기를 권장하고 **표를 한 어휘는 400개로 진로선택 및 전문교과 I 에 사용 하기를 권장한다.

학습자의 현재 수준(/) 설정

□ 영어과 교육과정: 어휘

□ 초등: 800+200
= 1,000

6. 생활 주변에서 흔히 사용하고 있는 아래의 낱말들은 외래어 가운데 학년별로 50단어씩 사용 가능하며 이는 새로운 어휘로 간주하지 않는다.

alarm, album, alcohol, amateur, ambulance, apartment, arch, bacon, badminton, bag, banana, belt, bench, biscuit, bonus, box, bus, butter, cabinet, cake, calcium, camera, camp, campaign, campus, card, carol, carpet, catalogue, center, champion, channel, chart, cheese, chicken, chocolate, click, coat, coffee, comic, computer, course, court, crayon, cream, cup, data, diamond, diet, disc, doughnut, drama, dress, drill, drum, echo, elevator, elite, energy, essay, event, fashion, feminist, fence, festival, fiction, film, fork, gallery, game, gas, golf, gown, graph, guard, guitar, gum, hamburger, harmony, highlight, hint, hormone, hotel, image, interior, internet, interview, issue, jacket, jam, jazz, juice, jump, kangaroo, kiss, kiwi, laser, league, lemon, lobby, magic, manual, marathon, market, mask, medal, media, melon, member, menu, message, model, motor, mystery, news, notebook, okay, opera, orange, oven, page, panda, parade, partner, party, pen, percent, piano, pie, pilot, pipe, pizza, plastic, plug, program, project, quiz, radio, recreation, rehearsal, ribbon, robot, rocket, rugby, salad, sample, sandwich, sauce, scarf, scenario, schedule, section, seminar, service, set, shirt, skate, sketch, ski, snack, soup, spaghetti, sponsor, sport, spray, spy, staff, star, steak, stereo, studio, style, sweater, tank, taxi, team, technique, technology, television, tennis, tent, terror, ticket, toast, tomato, topic, towel, track, truck, vaccine, veil, video, villa, violin, virus, vision, waiter, website, wine, yacht (200개)

학습자의 현재 수준(/) 설정

□ 영어과 교육과정: 어휘

▣ KICE Word Lister 2015

(주형미, 진경애, 김미경, 김성혜, 정채관, 박상복 2016)

정리

□ 학습자 수준에 적합한 다독/다청 자료란?

- ▣ 자료의 언어적 수준이 **학습자의 현재 수준(/)**보다 약간 높거나 낮음
- ▣ 자료에 제시된 외국어 입력의 90~98%(token 기준)가 **학습자가 이미 알고 있는 어휘로 구성**

정리

- 학습자 수준에 적합한 다독/다청 자료란?
 - ▣ 자료의 언어적 수준이 영어과 교육과정에 따른 학습자의 현재 수준보다 약간 높거나 낮음
 - ▣ 자료에 제시된 외국어 입력의 90~98%(token 기준)가 영어과 교육과정에 비추어 학습자가 이미 알고 있을 거라 예상되는 어휘

효과

→ 공교육과의 연계

선행연구

- 아동용 애니메이션 영화
(김기택과 성민창, 2019; 성민창과 김기택, 2021a)



선행 연구 결과

- 학교에서 교과서로 성실히 영어를 공부한, 졸업을 목전에 둔 초등학교 6학년 학생이 이해가 가능한 어휘?

→ 영어교육과정 어휘, 외래어, 고유명사, 감탄사 등

표 3. 애니메이션 영화별 초등 수준에서 이해 가능한 어휘의 구현 빈도 비율

분석 대상	이해 가능 어휘						이해 불가 어휘
	교육과정 어휘	외래어	단위 명칭	고유 명사	감탄사	합계	
1-Lion	5,879 (78%)	51 (1%)	13 (0%)	230 (3%)	315 (4%)	6,488 (86%)	1,067 (14%)
2-Goofy	4,353 (78%)	81 (1%)	29 (1%)	178 (3%)	190 (3%)	4,831 (86%)	773 (14%)
3-Notre	5,392 (76%)	72 (1%)	49 (1%)	303 (4%)	62 (1%)	5,878 (83%)	1,211 (17%)
16-Inside	6,856 (82%)	133 (2%)	48 (1%)	138 (2%)	185 (2%)	7,360 (88%)	1,020 (12%)
17-Dory	8,930 (80%)	176 (2%)	74 (1%)	410 (4%)	375 (3%)	9,965 (89%)	1,248 (11%)
18-Moana	5,213 (80%)	53 (1%)	13 (0%)	213 (3%)	86 (1%)	5,578 (86%)	936 (14%)
19-Zootopia	7,311 (74%)	96 (1%)	172 (2%)	279 (3%)	281 (3%)	8,139 (83%)	1,724 (17%)
	6,775 (79%)	77 (1%)	17 (0%)	463 (5%)	217 (3%)	7,549 (88%)	1,014 (12%)
	5,899 (78%)	91 (1%)	52 (1%)	237 (3%)	194 (3%)	6,472 (85%)	1,115 (15%)



표 4. 애니메이션 영화 내용 이해를 위해 듣기 전 활동에서 학습되어야 할 어휘의 수

분석 대상	구현 빈도 90% 기준	구현 빈도 95% 기준	구현 빈도 98% 기준
1-Lion	43	238	465
2-Goofy	30	168	867
3-Notre	87	318	530
4-Hercules	89	389	652
5-Mulan	89	299	465
6-Monsters	20	155	403
7-Ice	49	202	338
8-Incredibles	144	473	750
9-Madagascar	40	254	
10-Car	59	297	
11-Panda	37	181	
12-Tangled	33	201	
13-Toy	58	264	
14-Brave	69	209	
15-Frozen	12	132	
16-Inside	12	141	381
17-Dory	5	63	266
18-Moana	29	157	353
19-Zootopia	100	396	693
20-Coco	10	120	372
평균	51	233	478



발표 순서

- 연구의 동기
- 이론적 바탕
- 학습자 수준
- 아동문학작품 분석

표 1. 분석 대상: 14편의 아동 영문학 작품

연번	작품명	저자명	작품 개요
1	<i>A Single Shard</i>	Linda Sue Park	고려 시대 도자기 마을에 살던 소년 목이가 그토록 바라던 도공으로 성장하는 내용으로서, 고려청자라는 우리의 아름다운 예술과 역경을 딛고 일어서는 아동의 성장 스토리가 잘 어우러져 있다.
2	<i>Charlotte's Web</i>	E. B. White	소녀 Fern, 아기돼지 Wilbur, 거미 Charlotte의 우정 이야기로서, 이들의 진지한 소통과 서로를 위한 고귀한 희생이 감동적으로 다가온다.
3	<i>Dear Mr. Henshaw</i>	Beverly Cleary	소년 Leigh가 자신이 좋아하는 작가와 6년간 나누었던 편지들과 그가 쓴 일기 형식으로 구성된 소설로서, 이혼가정 아동의 고통과 외로움, 아파에 대한 그리움과 성장 이야기가 잘 녹아있다.
4	<i>Fantastic Mr. Fox</i>	Roald Dahl	여우이자 가장인 Fox씨가 그의 영리함과 친구들과의 협력을 통해 자신을 잡으려는 사악한 농부 3명의 계획을 통쾌하게 이겨내고서 가족, 친구들과 함께 살아갈 발판을 만드는 내용이다.
5	<i>Flora and Ulysses</i>	Kate DiCamillo	부모의 이혼 속에 냉소적으로 변한 소녀 Flora가 우연히 다람쥐 Ulysses의 목숨을 구해주고 서로 친구가 되면서, 시를 쓰는 이 다람쥐의 초능력으로 그녀와 갈등을 빚던 엄마와 화해하고 마음을 열면서 따뜻한 아이로 성장하는 내용이다.
6	<i>Frindle</i>	Andrew Clements	학생 Nick이 펜을 Frindle이라고 바꿔 부르면서 이 신조어의 사용을 막으려는 선생님의 의지를 꺾고 이 단어를 사전에까지 등재시키는 이야기로서, 교실 상황의 유쾌함과 아이의 창의성이 생생하게 나타난다.
7	<i>Holes</i>	Louis Sachar	우연히 누명을 쓰고 소년원으로 이송된 Stanley가 조상의 일과 자신의 현재 불운이 맞닿아 있음을 파악하고서 가족에 얹힌 저주를 용기 있게 풀어내는 아동소설이자 추리소설로서, 인종차별, 빈곤, 노숙 등의 사회문제까지 잘 다뤄낸 수작이다.
8	<i>Judy Moody Declares Independence</i>	Megan McDonald	기발하면서도 엉뚱한 소녀 Judy의 15권 시리즈 중 한 권으로, 미국 독립의 역사와 Judy의 발랄한 일상이 흥미롭게 어우러진다.
9	<i>Pippi Longstocking</i>	Astrid Lindgren	주근깨 많은 빨강머리 괴력 소녀 Pippi가 속물적인 어른들에게 발달함과 순수함으로 맞대응하면서, 잠깐의 이웃 친구들 및 동물들과 함께 하는 당찬 어린이의 일상을 유쾌하게 보여준다.
10	<i>The Little Prince</i>	Antoine de Saint-Exupéry	지구별로 온 어린 왕자가 사막에 불시착한 조종사와 친구가 되어 대화를 나누면서, 순수함 잃고 왜곡되어버리는 어른들의 세계를 비판적으로 고찰하고, 순수, 사랑, 인생에서 중요한 것들, 성장의 의미에 대한 교훈을 진지하게 담아낸다.
11	<i>The Magic Finger</i>	Roald Dahl	어른들의 부당함과 잔인함에 화가 났을 때 요술 손가락으로 어른들을 혼내고 잘못을 일깨워주며, 엉뚱하고 발랄하면서도 정의로운 한 소녀의 통쾌하고도 흥미진진한 환상 이야기이다.
12	<i>The Vanderbeekers of 141st Street</i>	Karina Yan Glaser	뉴욕 할렘을 배경으로, 집주인이 임대 계약 연장을 거부하자 세 입자인 가족들이 회의를 하며 이야기가 시작되는데, 집을 사주하려는 가족들의 필사적인 노력 속에 따뜻한 가족애가 돋보이는 이야기이다.

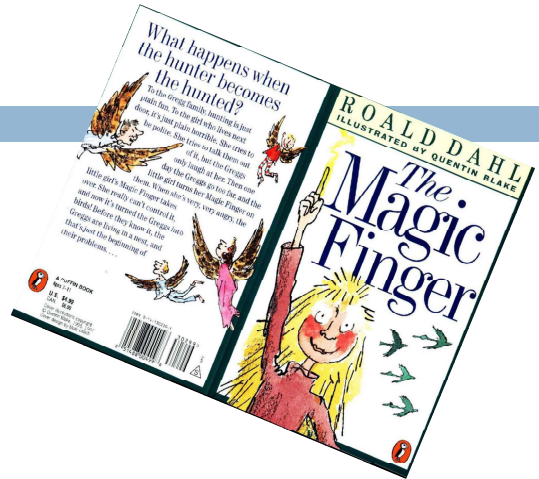


표 1. 계속

연번	작품명	저자명	작품 개요
13	<i>Where the Mountain Meets the Moon</i>	Grace Lin	중국 전통 민담에 기초한 환상 소설이자 영웅담으로서, 어린 소녀 Minli가 가족과 마을을 구하기 위해 떠나는 영웅적 여정을 중국 문화의 아름다운 배경 속에 다채롭고 환상적으로 그려낸다.
14	<i>Wonder</i>	R. J. Palacio	얼굴이 일그러진 장애티를 타고난 Auggie가 집에서 벗어나 학교에 다니기 시작하면서 불거진 학교에서의 괴롭힘을, 그가 가족들의 사랑과 친구들과의 우정으로 극복해 나가며 Auggie와 그의 가족 및 주변 친구들 모두 성장해가는 이야기이다. 장애, 집단 괴롭힘, 우정, 가족애, 성장의 주제가 각 인물들의 다각적인 시선을 통해 진지하면서도 흥미롭게 다뤄진다.

표 2. KICE Word Lister 2015 어휘 범주의 재구조화

범주	설명	예시	재구조화된 범주
초등권장	초등과정에서 사용하기를 권장하는 800개의 기본 어휘	have, way	초등 권장
초등권장의외	중등학교 일반과정에서 사용하기를 권장하는 1,800개의 기본 어휘	real, happen	중등 일반
고등 진로선택 및 전문교과 I	진로 선택 및 전문 교과에서 사용하기를 권장하는 400개의 기본 어휘	slight, thread	고교 전문/선택
감탄사·호칭/ 고유명사/ 외래어/수/ 알파벳·문자	감탄사, 호칭, 고유 명사, 기수, 서수, 흔히 사용되는 외래어, 알파벳 등 (새로운 어휘로 간주하지 않음)	oh, William, box, two	기타 이해가능
미확인어휘	교육과정에 수록되지 않은 어휘	basin, canteen bourgeoisie	미수록 고등 비영어 어휘

표 8. 각 작품의 어휘 수준별 구현 빈도

작 품	초등 수준 어휘(%)			상위 수준 어휘(%)			소계	합계(회)
	초등 권장	기타 이해가능	소계	중등 일반	고교 전문/선택	미수록 고등		
1	72.7	4.8	77.4	16.7	0.8	5.0	22.6	33,883
2	74.7	7.7	82.4	11.6	0.6	5.4	17.6	32,143
3	81.6	5.5	87.1	9.2	0.4	3.2	12.9	18,075
4	76.9	5.8	82.8	11.9	0.8	4.6	17.2	9,579
5	74.8	8.4	83.2	11.3	0.7	4.7	16.8	32,314
6	76.9	9.6	86.5	10.9	0.3	2.3	13.5	16,165
7	75.8	8.0	83.8	12.1	0.6	3.4	16.2	47,035
8	71.9	11.9	83.9	11.2	0.4	4.5	16.1	11,108
9	78.8	6.9	85.7	10.3	0.5	3.4	14.3	25,864
10	81.8	1.8	83.6	12.3	0.6	3.6	16.4	16,946
11	81.1	8.6	89.7	8.5	0.1	1.7	10.3	3,716
12	71.5	10.3	81.8	12.3	0.6	5.3	18.2	46,795
13	78.7	4.0	82.7	12.1	0.6	4.7	17.3	42,528
14	80.6	5.9	86.5	10.5	0.3	2.7	13.5	73,346

표 9. 작품별 이해를 위해 필요한 상위 수준 어휘 유형 규모

작품	90% 기준을 위한 추가 학습 어휘				95% 기준을 위한 추가 학습 어휘			
	중등 일반	고교 전문/선택	미수록 고등	합계	중등 일반	고교 전문/선택	미수록 고등	합계
11	1			1	21		1	22
3	15		1	16	111	6	15	132
6	23		1	24	175	1	11	187
14	26			26	207	4	28	239
8	30	1	4	35	141	3	35	179
9	49	1	3	53	215	11	39	265
4	52	1	6	59	166	10	30	206
10	50		12	62	210	12	30	252
7	53	2	8	63	255	14	36	305
5	70	4	9	83	257	14	66	337
13	78	2	9	89	291	12	50	353
2	83	3	19	105	324	17	59	400
12	183	6	17	206	560	32	137	729
1	191	6	22	219	476	30	91	597

What happens when the hunter becomes the hunted?

To the Gregg family, hunting is just plain fun. To the girl who lives next door, it's just plain horrible. She tries to be polite. She tries to talk them out of it, but the Greggs only laugh at her. Then one day the Greggs go too far, and the little girl turns her Magic Finger on them. When she's very, very angry, the little girl's Magic Finger takes over. She really can't control it, and now it's turned the Greggs into birds! Before they know it, the Greggs are living in a nest, and that's just the beginning of their problems....



A PUFFIN BOOK
Ages 7-11

U.S. \$4.99

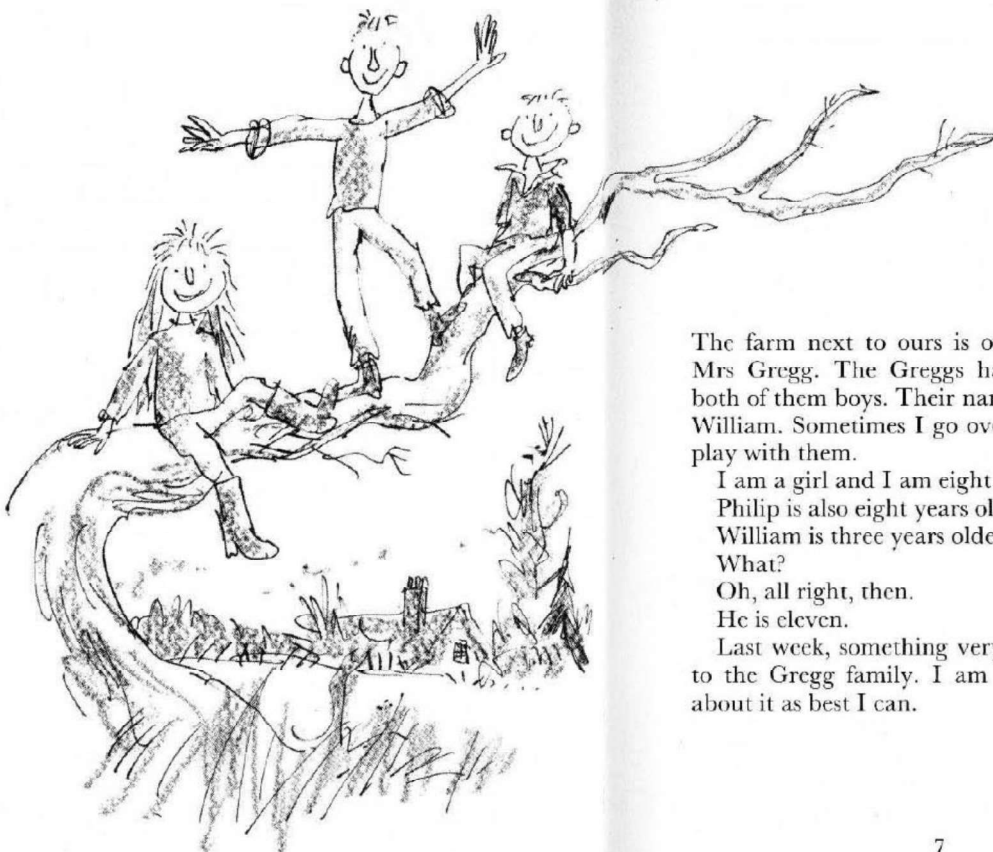
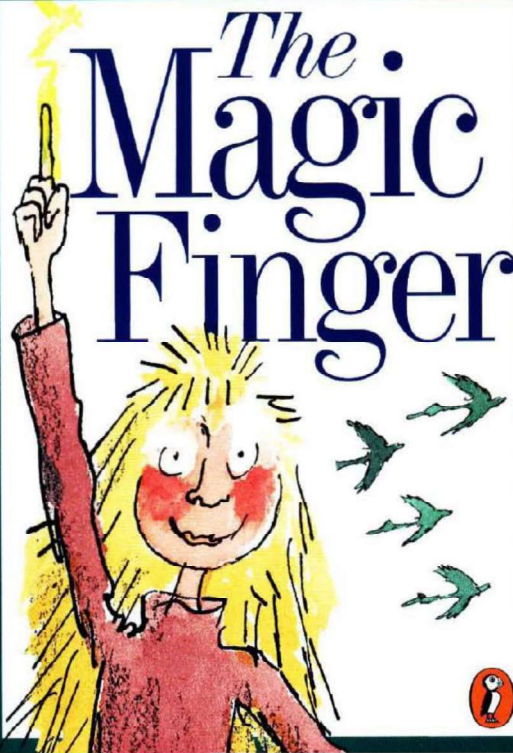
CAN. \$6.99

Cover illustrations copyright
© Quentin Blake, 1995, 1997
Cover design by Molly Leach

ISBN 0-14-130229-1



ROALD DAHL
ILLUSTRATED BY QUENTIN BLAKE



The farm next to ours is owned by Mr and Mrs Gregg. The Greggs have two children, both of them boys. Their names are Philip and William. Sometimes I go over to their farm to play with them.

I am a girl and I am eight years old.

Philip is also eight years old.

William is three years older. He is ten.

What?

Oh, all right, then.

He is eleven.

Last week, something very funny happened to the Gregg family. I am going to tell you about it as best I can.

- 정보의 홍수 시대: 필요한 정보 탐색의 중요성
- 학습자 수준에 적합한 다독/다청 자료를 찾는 방법론 하나를 제시

References (selected)

- 교육부. (2015). *영어과 교육과정*. 교육부.
- 김기택, 성민창. (2019). 초등영어 학습자 어휘 수준에 적합한 영어 애니메이션 영화 탐색. *영어학*, 19(4), 837-857.
- 김솔, 남혜림, 최수미, 차용석, 조규희, 이동환. (2017). 문학작품을 활용한 초등영어교육에 대한 교사들의 인식 연구. *초등영어교육*, 23(1), 55-83.
- 김은현, 박혜림. (2022). 다문화 아동문학을 활용한 초등영어와 다문화의 통합 교육: 예비교사 교육을 중심으로. *학습자중심교과교육연구*, 22(19), 177-198.
- 김혜리. (2009). 아동문학 활용 초등영어 교과교육 프로그램 내용 개발에 대한 연구. *외국어교육*, 16(2), 267-292.
- 박기화. (2015). 초등영어교수에서 장면 아동문학 작품의 활용. *학습자중심교과교육연구*, 5(6), 343-365.
- 성민창, 김기택. (2021a). 2015 개정 영어과 교육과정 초등 수준 어휘 목록을 활용한 영어 애니메이션 영화의 어휘 사용 분석. *영어학*, 21, 87-109.
- 성민창, 김기택. (2021b). 우리나라 초등영어교육에 적합한 팝송 탐색: 언어 수준과 내용 분석. *영어학*, 21, 450-471.
- 이병민. (2003). EFL 영어학습 환경에서 학습시간의 의미. *외국어교육*, 10(2), 107-129.
- Carver, R. P. (1994). Percentage of unknown vocabulary words in text as a function of the relative difficulty of the text: Implications for instruction. *Journal of Reading Behavior*, 26(4), 413-437.
- Cummins, J. (1981). Age on arrival and immigrant second language learning in Canada: A reassessment. *Applied Linguistics*, 11(2), 132-149.
- Day, R., & Bamford, J. (1998). *Extensive reading in the second language classroom*. Cambridge University Press.
- Hirsh, D. (1993). *The vocabulary demands and vocabulary learning opportunities in short novels*. Unpublished master thesis, Victoria University, Wellington, New Zealand.
- Hirsh, D., & Nation, I. S. P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language*, 8(2), 689-696.
- Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13(1), 403-430.
- Krashen, S. (1982). *Principles and practices in second language acquisition*. Pergamon.
- Krashen, S. (1993). The case for free voluntary reading. *Canadian Modern Language Review*, 50(1), 72-82.
- Krashen, S. (2004). *The power of reading: Insights from the research*. Heinemann.
- Krashen, S., & Terrell, T. (1983). *The natural approach: Language acquisition in the classroom*. Pergamon.
- Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines* (pp. 316-323). Multilingual Matters.
- Long, M. H. (1983). Native speaker/non-native speaker conversation and the negotiation of comprehensible input. *Applied Linguistics*, 4(2), 126-141.
- Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, 63(1), 59-82.
- Nation, I. S. P. (2013). Vocabulary size in a second language. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 6674-6677). Wiley-Blackwell.
- Nation, I. S. P., & Hwang, K. (1995). Where would general service vocabulary stop and special purposes vocabulary begin? *System*, 23, 35-41.
- Snow, C., & Ferguson, C. (1977). *Talking to children: Language input and acquisition*. Cambridge University Press.

한국어능력시험 II 오답지의 화용적 오류 유인 요소 분석

- 간접 말하기 문항과 간접 행동 문항을 대상으로

정윤희
석사과정 학생, 고려대학교 언어학과

Introduction

Background

Method

Analysis

Conclusion

Limitation

연구의 목적과 필요성

- 맥락적 층위에서 한국어능력시험 II 오답지 분석
- 한국어능력시험은 한국어 학습자의 의사소통 능력을 실질적으로 평가하는 데 한계에 봉착
- 한국어능력시험 평가 체계 개선을 위해 문항 출제 방식 개선과 출제 및 채점 전문가를 지속적으로 양성할 계획 추진(교육부, 2022)
- 한국어능력시험 II에서 의사소통 능력과 직결되는 문항 유형의 오답지를 대상으로 화용적으로 어떠한 오류를 내포하고 있는지 분석
- 한국어능력시험에서 KSL 또는 KFL 학습자의 의사소통 역량이 강조되는 만큼 학습자의 취약한 화용적 능력을 파악하여 보완하는 데 기여하고자 함

2

연구의 배경

- 기존의 한국어능력시험 연구는 듣기와 읽기 영역에서 의사소통을 실질적으로 평가하고 있는지에 대한 텍스트 분석이 주로 이루어졌고, 오답 연구는 이루어지지 않았음
- 한국어능력시험의 오답 연구는 전무
- 오답 유형 또한 단어의 형태나 의미와 관련한 유형에 한정됨. 맥락과 관련한 오답 유형은 주제/화제(어)와 관련한 단어를 사용하는 정도에 그침 (Goodrich, 1977)
- 언어 표준화 시험이 수험자들의 화용적 지식을 측정하기 위해 듣기 문항을 출제하려고 노력하고 있으나, 실질적으로 측정하고 있는지에 대한 연구는 거의 이루어지지 않고 있음(Allami & Aghajari, 2014)
- 선다형 문항 제작 시 규칙이나 가이드라인은 있으나 그럴듯한(plausible) 오답 제작 참고지표(reference)가 없으므로 맥락 층위의 오답 연구가 시급함(Hoshino, 2013)

3

AI에 기반한 맥락적 층위의 자동 오답 생성

Zhou, 2020

- 문법과 의미 층위의 자동 오답 생성 연구만 이루어짐
- 오답지가 지문의 주제와 관련되지 않거나 지문의 맥락과 일관적이지 않는 문제점이 발생한다.
- 빈칸 채우기(Liang et al, 2017) 또는 cloze MCQ에서 단어 수준 또는 구문 수준 오답지에서 탈피하고 긴 오답지(long distractor)를 생성해야 한다.

4

수능 영어 오답지의 화용적 오류 유인요소 분석 (신유경 · 유석훈, 2021)

M: Honey, I've just left work. I'll be home in half an hour.

W: Good. Is it possible for you to stop by the dry cleaner's shop and pick up my dress?

M: Sure. Can you tell me where the shop is located?

W: _____.

- ① Okay. I'll send the address to your phone. (정답)
- ② Yes. I'll have your dress cleaned by noon
- ③ Of course. I'll open the shop tomorrow. (동일 지시체; Same Referent)
- ④ No. I'm not moving to a new place.
- ⑤ Too late. I'm already back at home. (상반 내용; Opposite)

화용적 오류 유인요소 Pragmatic Error Triggers

형태와 어휘를 기반으로 하면 오답이 아니지만, 맥락을 기반으로 하면 오답이 되도록 만드는 요소

(신유경 · 유석훈, 2021)

5

간접 말하기 문항



다음 대화를 잘 듣고 이어질 수 있는 말을 고르십시오. (각 2점)

남자: 왜 그래요? 뭐 잃어버렸어요?

여자: 제 가방이 어디에 있는지 모르겠어요.

남자: _____. (52회 4번)

- 1. 찾아서 다행이네요. (다른 내용 + 중요 표현)
- 2. 천천히 찾아봐요. (정답)
- 3. 가방을 찾고 있는데요. (잉여 중복 + 중요 표현 + 차례오류)
- 4. 아무리 찾아도 없어서요. (차례 오류)

6

간접 행동 문항



다음 대화를 잘 듣고 여자가 이어서 할 행동으로 알맞은 것을 고르십시오. (각 2점)

남자: 누나, 벽시계가 안 가는 것 같은데. 건전지가 다 됐나 봐.

여자: 어, 그러네. 건전지가 어디 있었던 것 같은데.....

남자: 안방 서랍 안에 몇 개 있을 거야. 가져올게.

여자: 아니야. 내가 찾아올 테니까 너는 시계 좀 내려 줘. (60회 11번)

- ① 건전지를 가지러 간다. (정답)
- ② 현재 시간을 확인한다.
- ③ 시계를 벽에서 내린다.
- ④ 건전지를 서랍에 넣는다.

7

연구 문제



01

한국어능력시험 II 오답지의 화용적 오류 유인요소(Pragmatic Error Triggers; PET)는 무엇인가?

02

한국어능력시험 간접 말하기 문항 오답지의 PET 빈도는 무엇인가?

03

한국어능력시험 간접 행동 문항 오답지의 PET 빈도는 무엇인가?

8

Grice's maxim (1975)

- 협조의 원칙(Cooperative Principle) : 대화자가 참여하고 있는 대화의 각 진행 단계에서 대화의 목적이나 방향에 요구되는 것만큼 기여해야 한다(Grice 1975: 45).
- 협조의 원칙 중 방법의 격률은 발화를 명료하게 함으로써 L2 이해가 낮은 학습자도 L2 담화를 이해할 수 있고, L2 습득을 향상시킬 수 있다(She, 2022).
- 간접 말하기 문항과 간접 행동 문항은 각각 3회 교대, 4회 교대로 많은 맥락적 정보를 포함하고 있지 않으므로 한국의 사회적, 문화적 요인을 담아 함축적 의미를 파악할 수 있는 맥락을 구성하는 데 어려움이 따른다.

9

연구 방법

화용적 오류 유인요소 유형화

1. 초안 작성 후 전문가와의 논의로 화용적 오류 유인요소 유형화
2. 평가자와 태깅을 수행하면서 화용적 오류 유인요소 합의
3. 연구자 간 태깅 결과 일치도와 연구자와 평가자 간 태깅 결과 일치도가 Cohen's kappa 0.8 넘을 때까지 합의과정 반복

데이터 분석

1. 화용적 오류 유인요소의 절대빈도, 비율
2. SPSS로 언어 간, 문항 간 PET 빈도의 피서정확검정 또는 카이제곱검정 수행

10

연구 대상

한국어와 수능
영어의 PET 비
교

구분	회차	3회 교대 간접 말하기 문항	문항 당 오답 지 수	총 오답지	
한국어능력시험 II	8회	5문항	3개	120개	
수능 영어	9회	2문항	4개	72	120개
영어 모평	6회(/14회)	2문항	4개	48	

한국어능력시
험 간접 말하기
& 간접 행동
문항 PET 비교

구분	회차	문항 수	문항 당 오답지 수	총 오답지
간접 말하기 문항	6회	5문항	3개	96개
간접 행동 문항	8회	4문항		

11

한국어의 화용적 오류 유인요소 (Pragmatic Error Triggers)

Grice' maxims		Pragmatic Error Triggers	
Maxim of Quantity	내용 초과 오류	잉여 중복	Redundancy(RE)
		초점 오류	Wrong Focus(WF)
	내용 부족 오류	모호한 대명사/관형사	Vague Pronoun/Determiner(VP/VD)
Maxim of Quality	내용 불일치 오류	다른 내용	Different Contents(DC)
Maxim of Relevance	내용 관련성 오류	주제 관련	Topic-Relevant(TR)
		주제 비관련	Topic-Irrelevant(TI)
Maxim of Manner	순서 오류	순서 오류	Order(OR)
		차례 오류	Turn(TU)

12

Grice's maxim (1975)

Maxim of Quantity

- ① 당신이 기여하는 뜻은 필요한 만큼 제보적(informative)이어야 한다.
- ② 당신이 기여하는 뜻은 필요 이상으로 정보를 주어서는 안 된다.

- 잉여 중복(Redundancy; RE) : 대화의 표현/내용과 같거나 유사한 것의 잉여/과잉이 되는 중복
- 초점 오류(Wrong Focus; WF) : 제 3자뿐만 아니라 남, 여 외 타인이 할 수 있는 말로 범위 확대
- 모호한 대명사/관형사(Vague Pronoun/Determiner; VP/VD) : 대화에서 근거/단서를 찾을 수 없는 대명사나 관형사 사용

13

양의 격률 하위 유형

- 잉여 중복(Redundancy)

다음 대화를 잘 듣고 이어질 수 있는 말을 고르십시오. (각 2점)

남자: 왜 그래요? 뭐 잃어버렸어요?

여자: 제 가방이 어디에 있는지 모르겠어요.

남자: _____ (52회 4번)

1. 찾아서 다행이네요.
2. 천천히 잘 찾아봐요. (정답)
3. 가방을 찾고 있는데요. (RE)
4. 아무리 찾아도 없어서요.

14

양의 격률 하위 유형

- 초점 오류(Wrong Focus)

다음 대화를 잘 듣고 이어질 수 있는 말을 고르십시오. (각 2점)

남자: 지영 씨, 혹시 친구 중에 아르바이트할 만한 사람이 있을까요?

여자: 글세요. 잘 모르겠는데 친구들한테 한번 물어볼까요?

남자: _____ (47회 6번)

1. 네. 일할 곳을 찾고 있어요. (WF)
2. 네. 제가 대신 해 드릴게요.
3. 네. 좀 알아봐 주면 좋겠어요. (정답)
4. 네. 그 친구는 아르바이트해요.

15

Grice's maxim (1975)

Maxim of Quality

- ① 거짓이라고 믿는 것은 말하면 안 된다.
- ② 충분한 증거가 부족한 것은 말하면 안 된다.

- 다른 내용(Different Content; DC) : 대화의 상황/내용, 해당 차례에 할 수 있는 말 또는 맥락의 전제와 다름

여자: 잠깐 약국 좀 다녀올게요. 두통약 좀 먹어야겠어요.

남자: 몸이 안 좋으면 빨리 집에 가서 쉬세요. 나머지 전시회 상품은 제가 정리할 테니까요.

여자: 아니에요. 아직 설명 자료도 못 만들었는데요.

남자: 네, 그럼 다녀오세요. 이걸 제가 정리하고 있을게요. (47회 10번)

- ① 집에 간다. (DC)

16

Grice's maxim (1975)

Maxim of Relevance

관련성 있게 말해야 한다.

- 주제 관련(Topic-Relevant; TR) : 주제/중심 소재와 관련된 중요 표현
- 주제 비관련(Topic-Irrelevant; TI) : 관련되지 않은 단어, 구, 절을 사용(이 단어, 구, 절만 바꾸면 맥락에 부합)

여자 : 저, 갑자기 비가 와서 그러는데 요즘도 **우산** 빌려 주세요?

남자 : 어쨌죠? **우산 대여 서비스**가 지난달부터 중단되었습니다. 우산을 반납하는 분들이 너무 적어서 서비스를 계속하기가 힘들어서요.

여자 :

1. 신청서만 쓰면 빌릴 수 있어서 참 편해요. (TR)
2. **빌려가는** 사람들이 생각보다 적었나 봐요. (TR + TI)
3. 우산은 내일 출근하는 길에 갖다 드릴게요. (TR)
4. 필요할 때 빌릴 수 있어서 좋았는데 아쉽네요. (정답)

17

Grice's maxim (1975)

Maxim of Manner

- ① 애매함을 피해야 한다.
- ② 간결하게 말해야 한다.
- ③ 순서대로 말해야 한다.

- 순서 오류(Order; OR) : 해당 발화 이전 또는 이후에 할 수 있는 말
- 차례 오류(Turn; TU) : 남자와 여자의 말차례 오류

18

관련성 격률 하위 유형

- 순서 오류(Order; OR), 차례 오류(Turn; TU)

다음 대화를 잘 듣고 여자가 이어서 할 행동으로 알맞은 것을 고르십시오. (각 2점)

남자: 누나, 벽시계가 안 가는 것 같은데. 건전지가 다 됐나 봐.

여자: 어, 그러네. 건전지가 어디 있었던 것 같은데.....

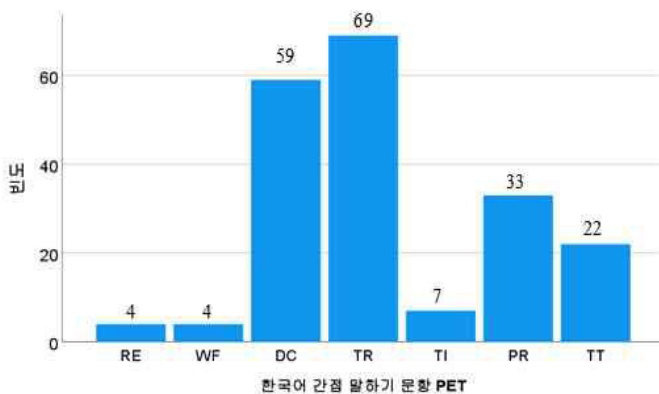
남자: 안방 서랍 안에 몇 개 있을 거야. 가져올게.

여자: 아니야. 내가 찾아올 테니까 너는 시계 좀 내려 줘. (60회 11번)

1. 건전지를 가지러 간다. (정답)
2. 현재 시간을 확인한다. (OR + TU)
3. 시계를 벽에서 내린다. (OR + TU)
4. 건전지를 서랍에 넣는다.

19

TOPIK II 간접 말하기 문항 오답지의 PET 빈도



화용적 오류 유인요소	계	Grice 격률에 따른 빈도 합계
잉여 중복(Redundancy; RE)	4	8
초점 오류(Wrong Focus; WF)	4	
모호한 대명사/관형사(Vague Pronoun/Determiner; VP/VD)	0	
다른 내용(Different Contents; DC)	59	59
주제 관련(Topic-Relevant; TR)	69	76
주제 비관련(Topic-Irrelevant; TI)	7	
순서 오류(Order; OR)	33	55
차례 오류(Turn; TU)	22	
계	198	198

20

TOPIK II와 수능 영어의 화용적 오류 유인요소
- PET 정의가 확대된 유형

한국어	영어
초점 오류(Wrong Focus)	
대화의 초점이 해당 청/화자에서 제3자로 잘못 이동	
대화에 언급된 + 대화에 언급되지 않은 제3자	대화에 언급된 제3자

한국어	영어
모호한 대명사/관형사 (Vague Pronoun/Determiner)	모호한 대명사 (Vague Pronoun)
대화에서 근거를 찾을 수 없는 대명사 사용	
대명사와 관형사 모두 태깅	대명사에 한정해 태깅

21

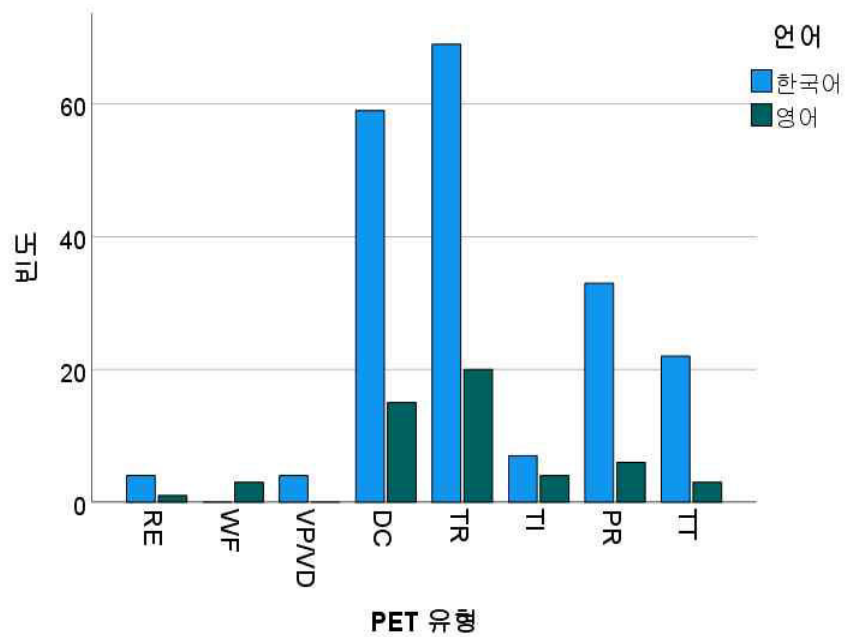
TOPIK II와 수능 영어의 화용적 오류 유인요소
- PET 정의가 확대된 유형

한국어	영어
주제 관련 (Topic-Relevant)	동일 지시체 (Same Referent)
대화의 주제/중심 소재 사용	
오답지의 단어, 구, 절, 표현 태깅	오답지의 지시체 태깅

한국어	영어
주제 비관련 (Topic-Irrelevant)	틀린 지시체 (Topic-Irrelevant)
잘못된 대상을 지칭하여 대화의 맥락에서 벗어남	
오답지의 단어, 구, 절 태깅	오답지의 지시체 태깅

22

TOPIK, 수능 영어의 간접 말하기 문항 PET 빈도



23

TOPIK, 수능 영어의 간접 말하기 문항 PET 빈도

구분				한국어 PET		영어 PET		$\chi^2(p)$
				빈도	상대빈도	빈도	상대빈도	
동일한 PET 정의	1	한국어	RE	4	2.0	1	1.9	1.000
		영어						
	2	한국어	DC	59	29.8	15	28.8	0.009 ^a (0.927)
		영어	OP					
	3	한국어	OR	33	16.7	6	11.5	0.777 ^a (0.378)
		영어						
	4	한국어	TU	22	11.1	3	5.8	0.311
		영어						
확장된 PET 정의	5	한국어	WF	4	2.0	0	0	0.584
		영어						
	6	한국어	VP/VD	0	0	3	5.8	0.008**
		영어	VP					
	7	한국어	TR	69	34.8	20	38.5	0.284 ^a (0.594)
		영어	SR					
	8	한국어	TI	7	3.5	4	7.7	0.245
		영어	WR					
전체				198	100%	52	100%	

* $p < .05$, ** $p < .01$, *** $p < .001$

24

절대적 빈도에 차이가 나는 경우

- 화용적 오류 유인요소 정의 확대되었다(Redundancy, Different Contents, Order, Turn 제외)
- 한국어는 고맥락 의사소통(high-context communication)으로 정보가 맥락 의존적이며 비명시적인 반면, 영어는 저맥락 의사소통(low-context communication)으로 정반대의 경향을 보인다(Hall, 1981).
- 한국어는 주어와 서술어의 위치가 유동적으로 바꿀 수 있기 때문에 문맥 정보가 중요한 역할을 한다.
- 영어의 문법 구조는 상대적으로 고정적이며, 주어와 서술어의 위치도 고정적이다.

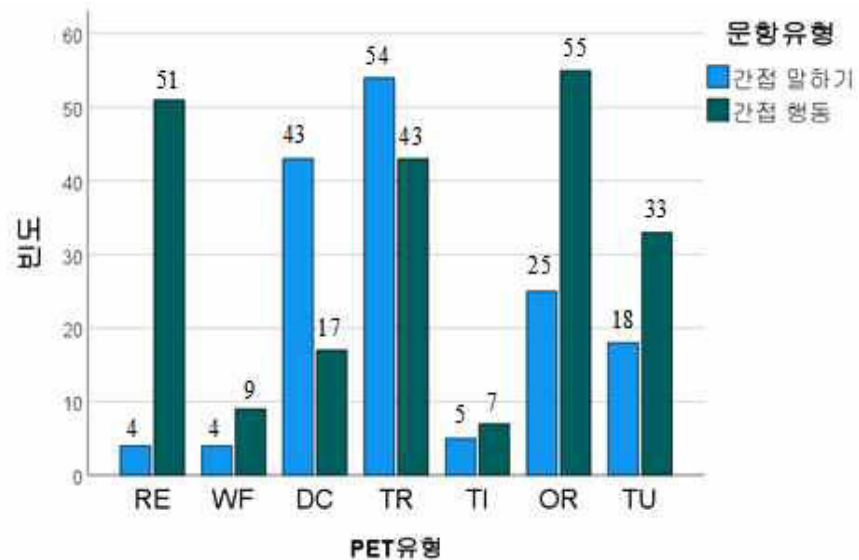
25

상대적 빈도가 비슷한 이유

- 시험 출제자들의 오답 구성 패턴이 비슷한 것을 확인
- 주제 관련(Topic-Relevant; TR)은 대화의 주제/중심소재가 되므로 맥락과의 연관성이 있어 보이도록 학습자들을 혼동시키기에 좋고, 비교적 용이하게 만들 수 있는 유형이므로 가장 높은 빈도를 보임
- 3회 교대의 맥락에서 Vague Determiner/Pronoun를 활용하여 오답지를 구성하는데 맥락정보의 한계를 보임

26

TOPIK II 간접
말하기 문항,
간접 행동 문항
PET 빈도



27

TOPIK II 간접
말하기 문항,
간접 행동 문항
PET 빈도

N = 368

구분		간접 말하기 PET		간접 행동 PET		x ² (p)
		빈도	상대빈도	빈도	상대빈도	
1	RE	4	2.6	51	23.7	46.624 ^a (0.000***)
2	VP/VD	0	0	0	0	1.000
3	WF	4	2.6	9	4.2	0.648 ^a (0.421*)
4	DC	43	28.1	17	7.9	26.722 ^a (0.000***)
5	TR	54	35.3	43	20.0	10.772 ^a (0.001***)
6	TI	5	3.3	7	3.3	1.000
7	OR	25	16.3	55	25.6	4.084 ^a (0.043*)
8	TU	18	11.8	33	15.3	0.962 ^a (0.327*)
전체		153	100%	215	100%	

* $p < .05$, ** $p < .01$, *** $p < .001$

TOPIK II PET 2개 이상 조합 빈도

간접 말하기 문항

구분	TR+DC	TR+OR	TR+TU	OR+TU	DC+OR
빈도 순위	1	2	3	4	5
빈도	20	15	7	7	4

간접 행동 문항

구분	RE+OR	RE+TU	TR+OR	OR+TU	RE+TR
빈도 순위	1	2	3	4	5
빈도	25	24	21	18	18

29

분석 결과

- 간접 말하기 문항에서 가장 많이 나타났던 PET 유형은 주제 관련(Topic-Relevant; TR)이었던 반면, 간접 행동 문항에서는 순서 오류(Order; OR)이다. 간접 말하기 문항에서는 대화의 핵심 소재/화제, 주제를 활용하여 오답지를 구성한 경우가 많은 반면, 간접 행동 문항에서는 시간적 순서를 활용하여 오답지를 구성하는 경우가 많았다.
- 간접 행동 문항은 문항 특성상 대화문에 제시된 이어질 행동의 순서들을 차례로 파악하는 문항이므로 OR을 활용한 오답지 구성이 가장 많은 빈도를 보였다.
- PET 유형 중 가장 높은 유의성을 가진 유형은 p-value가 0.000***인 RE, DC, 0.001***인 TR이다. 간접 행동 문항 오답지에서는 대화에서 언급한 발화 및 행동을 그대로 제시하여 학습자는 어떤 행동을 우선적으로 대화에 이어서 할 것인지 확인해야 한다.
- 모호한 대명사/관형사(Vague Pronoun/Determiner) 유형은 두 문항 유형에서 한 번도 나타나지 않았다.

결론

- 격률에 따른 PET 유형화는 잉여 중복, 초점 오류, 모호한 대명사/관형사, 다른 내용, 주제 관련, 주제 비관련, 순서 오류, 차례 오류로 나타났다.
- 한국어와 영어의 PET 빈도 차이를 확인한 결과, 절대 빈도에는 차이가 있고 상대 빈도 순위에는 차이가 없다. 모호한 대명사/관형사 유형을 제외하고 나머지 PET 유형에서 유의한 차이가 없는 것으로 확인되었다.
- 한국어능력시험 II 내 간접 말하기 문항과 간접 행동 문항의 문항간 6개 PET 유형에서 유의한 빈도 차이를 보였다. 대화의 순서를 파악하고자 하는 간접 행동 문항의 특성으로 인해 순서 오류의 빈도가 가장 많았다.

31

연구의 제한점 & 의의

연구의 제한점

1. 적은 수의 간접 말하기 문항, 간접 행동 문항 수
2. 화용적 오류 유인요소에 따른 응답률을 비교하지 못함. 학습자가 취약한 화용적 오류 유인요소를 파악하지 못했음
3. 총 단어 및 문장 수, 지문 당 평균 길이, 어휘 난이도, 통사적 복잡성과 같은 언어적 요소를 다면적으로 고려하지 못하였음

의의

1. 맥락적 층위에서 한국어의 응답 분석 시도
2. 메타인지를 지닌 KSL 또는 KFL 학습자의 취약한 화용적 능력 파악
3. 화용적 능력을 측정하는 문항 응답지에 한해 객관적인 응답 제작에 참고

32

감사합니다

정윤희 jyunhi12@naver.com

Corpus Linguistics and Digital Humanities

장세은 (한국해양대학교)

한국코퍼스언어학회·한국영어학회 여름 공동학술대회(2023. 6. 3. 고려대학교)

Contents

1. Definitions and research fields of corpus linguistics and digital humanities

- Corpus Linguistics (CL) vs. Digital Humanities (DH)
- A brief timeline of the evolution of CL and DH

2. Intersection between CL and DH and significant roles that CL has played in DH

- Previous studies vs. ChatGPT and BARD

3. Trends in interest over time about the comparison of CL and DH

- **Language visualisation**, Textual analysis, AI machine translation, Language network analysis, Sentiment analysis, Topic modeling analysis

4. Visual representations in linguistic theories

- Sounds/Words/Syntax/Semantics/Discourse

5. Case studies

- Textual analysis / AI machine translation
- Collocation analysis / Keyword analysis
- Language network analysis
- Sentiment analysis
- Topic modeling analysis

6. Summary and future directions

1. Definitions and research fields of corpus linguistics and digital humanities

1.1. Corpus Linguistics (CL)

- A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research (Sinclair, 2004).
- CL serves to answer two fundamental research questions:
 - (1) What particular patterns are associated with lexical and grammatical features?
 - (2) How do these patterns differ within varieties and registers?
- Corpus-based research in linguistics (Meyer, 2004)
 - Grammatical studies of specific linguistic constructions
 - Reference grammars
 - Lexicography
 - Language variation
 - Historical linguistics
 - Contrastive analysis and translation theory
 - Natural language processing
 - Language acquisition
 - Language pedagogy

1.2. Digital Humanities (DH)

What is digital humanities?

→ Application of techniques from computer science to matters from humanities



Digital humanities combines the methodologies from traditional humanities disciplines (such as history, philosophy, linguistics, literature, art, archaeology, music, pedagogy, and cultural studies) and social sciences, with tools provided by computing (such as Hypertext, Hypermedia, **data visualisation**, information retrieval, **data mining**, **statistics**, **text mining**, digital mapping, **topic modeling**, etc.), and digital publishing.

→ The interaction of computing or digital technologies and the disciplines of the humanities.

https://en.wikipedia.org/wiki/Digital_humanities

※ Computational Linguistics and Literature

Key words: **Corpus analysis** of literature, story analysis, computational analysis of poetry, **machine translation** of literature, **sentiment analysis** in poetry, ancient text decipherment and reconstruction

※ Definition and research fields of DH answered by ChatGPT 4.0

Definition: Digital humanities is an interdisciplinary field of study that combines computational methods with the disciplines of the humanities. This includes but is not limited to history, literature, philosophy, art history, musicology, cultural studies, linguistics, and other fields within the humanities. In digital humanities, computational tools and methods (such as [data visualization](#), [data mining](#), [machine learning](#), [text analysis](#), and digital mapping) are used to generate new insights into humanistic questions, to create digital archives and resources, to critically assess the impact of digital technologies on culture and society, and to explore new forms of teaching and scholarly communication in the digital age.

Research fields

- [Text Analysis](#) involves using computational techniques to analyze and visualize patterns in texts. Methods include [text mining](#), [topic modelling](#), [sentiment analysis](#), and stylometry.
- [Digital Pedagogy](#) involves the use of digital tools in teaching and learning, as well as the development of new pedagogical approaches suitable for the digital age.
- [Critical Code and Software Studies](#) involves the study of software and code as cultural artifacts, which includes understanding the cultural, social, and political implications of software and algorithms.
- Other research fields: Digital Archives and Libraries / Digital Scholarly Editing and Publishing / Digital Mapping and Spatial Humanities / Cultural Analytics / Public Humanities and Citizen Science / Virtual and Augmented Reality / Internet and Media Studies.

1.3 A brief timeline of the evolution of CL and DH

Here's a brief timeline that highlights some key developments in the evolution of both corpus linguistics and digital humanities:

Corpus Linguistics	Period	Digital Humanities	AI
- The completion of the Electronic Numerical Integrator and Computer (ENIAC, 1946) ¹⁾	1940s	Roberto Busa's Index Thomisticus (1946~1980) ²⁾	- Artificial Neural Network Model by Warren McCulloch & Walter Pitts (1943)
- The birth of machine translation in mid-1950s due to the Cold War - Noam Chomsky's syntactic structures (1957) - First concordance programs (1958) --> Busa's work working with IBM	1950s	- Early computational linguistics work (machine translation projects from 50s to 60s) - Zipf's Law ³⁾	- Alan Turing (Turing test : "Machine can think") ⁴⁾ - AI term coined by John McCarthy (1956) - Perceptron by Frank Rosenblatt (1957) algorithm as an origin of deep neural network (deep learning) - MIT Computer Science and Artificial Intelligence Laboratory (CSAIL) (M. L. Minsky, 1959), not officially

<ul style="list-style-type: none"> - First-generation corpus: Brown Corpus (1961) - John Sinclair's pioneer work from 1965 at the Uni.of Birmingham to 2000 	1960s	<ul style="list-style-type: none"> - The Association for Computational Linguistics (ACL, 1964) 	<ul style="list-style-type: none"> - Stanford Artificial Intelligence Laboratory (SAIL) founded by John McCarthy (1963)
<ul style="list-style-type: none"> - Lancaster-Oslo/Bergen (LOB) - London-Lund Corpus (first Spoken English) 	1970s	<ul style="list-style-type: none"> - Expansion of Humanities Computing - Development of Text Encoding Standards - Databases and Hypertext (--> World Wide Web) 	<ul style="list-style-type: none"> - Logic Machine (or Logic Theorist) based AI research (1972) - First AI Winter (Mid 70s ~ early 80s) (The Lighthill Report --> a critical review of AI research in the UK, 1973)
<ul style="list-style-type: none"> - Second-generation corpus: Birmingham Collection of English Text (18million words) 	1980s	<ul style="list-style-type: none"> - Advent of personal computers - <i>Digital Scholarship in the Humanities</i> (1986) SSCI, A&HCI, Scopus - Text Encoding Initiative (TEI, 1989)⁵⁾ 	<ul style="list-style-type: none"> - AI's Boom times (1980s) - Second AI Winter (1988, late 80s ~ early 90s)
<ul style="list-style-type: none"> - Third-generation corpus: BNC (1991) - COBUILD English language dictionary (1993) - Translation studies (1993) - WordSmith Tools (1996) - <i>International Journal of Corpus Linguistics</i> (1996) Scopus - Wmatrix (USAS and CLAWS, 1996) - Biber et al. (1999) <i>Longman grammar and spoken and written English</i> 	1990s	<ul style="list-style-type: none"> - WWW and the Explosion of Digital Content - Early Digital Libraries and Archives - A shift from Humanities Computing to Digital Humanities - TACT (Textual Analysis Computing Tools) developed by the Univ. of Toronto 	<ul style="list-style-type: none"> - The mid to late 1990s : a period of rejuvenation for AI - Emergence of the Internet (1995) - Development of Support Vector Machines (SVMs, 1997) - Rise of Machine Learning & Emergence of AI in E-commerce with the launch of companies like Amazon and Netflix in late 1990s
<ul style="list-style-type: none"> - AntConc (2002) - Sketch Engine (2004) - 21st-Century corpus: WebCorp project (e.g. Corpus of Contemporary American English (COCA, 2006) 	2000s	<ul style="list-style-type: none"> - Alliance of Digital Humanities Organizations (ADHO, 2004) - <i>Digital Humanities Quarterly</i> (2007), ESCI 	<ul style="list-style-type: none"> - The first autonomous robotic car by Carnegie Mellon Univ. (2001) - Facebook founded (2004) - Deep Learning (Geoffrey Hinton, 2006) - Google's self-driving car project (2009) - Word2Vec and Word Embedding Models (2009)
<ul style="list-style-type: none"> - Expansion of Web as a Corpus (NOW, GloWbE) & Large Scale Linguistic Corpora (COCA crossed the one-billion word mark in the mid-2010s) - NLP and Machine Learning 	2010s	<ul style="list-style-type: none"> - Department of Digital Humanities, King's College London (2011) BA/MA/PhD - <i>Distant Reading</i> by Prof. Franco Moretti at Stanford Univ. (2011) 	<ul style="list-style-type: none"> - Deep Learning-based AI Companies: (Google, Facebook 2013, Baidu 2014)

- COVID-19 corpora (e.g., Coronavirus, COVID-19 English Twitter Dataset)	2020s	<ul style="list-style-type: none"> - Response to COVID-19 - Growth in AI and Machine Learning Applications - Increased Collaboration Across Disciplines 	ChatGPT 3.5 (2022.11.30.) ChatGPT 4.0 (2023.03.14.) by OpenAI Bard (2023.05.10.) by Google
--	-------	--	--

- 1) The completion of the Electronic Numerical Integrator and Computer (ENIAC), one of the earliest electronic general-purpose computers, laid the groundwork for the computational analysis of text that would become central to corpus linguistics.
- 2) The Index Thomisticus was a digital humanities project begun in the 1940s (with IBM in 1949) that created a concordance to 179 texts centering around [Thomas Aquinas](#). Led by Roberto Busa, the project indexed 10,631,980 words over the course of 34 years, initially onto punched cards. It is considered [a pioneering project in the field of digital humanities](#).
- 3) George Kingsley Zipf publishes his seminal work *The Psycho-Biology of Language*, which introduces the concept of the Zipf distribution, a mathematical law that describes the [frequency of words in a language](#).
- 4) Turing, Alan (1950) "Computing Machinery and Intelligence." In this paper, Turing proposed the [Turing test](#), a test of a machine's ability to exhibit intelligent behavior equivalent to, or indistinguishable from, that of a human. The Turing test is still used today as a benchmark for artificial intelligence.
- 5) Founded to develop guidelines for encoding machine-readable texts in the humanities, the Text Encoding Initiative (TEI) has since incorporated many concepts from corpus linguistics to standardize the digital representation of texts for online research and teaching.

2. Intersection between CL and DH and significant roles that CL has played in DH

2.1 Previous studies

- (1) Jensen, K. E. (2014). Linguistics in the digital humanities: (computational) corpus linguistics. *MedieKultur: Journal of Media and Communication Research* 30(57), 115.-132.

This article provides an overview of the main principles of corpus linguistics and the role of computer technology in relation to data and method and also offers a bird's-eye view of the history of corpus linguistics with a focus on its intimate relationship with digital technology and how digital technology has impacted the very core of corpus linguistics and shaped the identity of the corpus linguist.

- (2) Zottola, A. (2020). Corpus linguistics and digital humanities Intersecting Paths: A Case Study from Twitter. *América Crítica* 4(2), 131-141.

- The purpose: the critical discussion about the role of CL within the field of DH.
- A case study as an example of how the two approaches can come together within the framework of American Studies.
- Data and methodology:

- Scraping Twitter in 6 months from October 2018 to March 2019 and compiling a corpus of 438,723 tokens.
- Watrix for generating key semantic domains against BNC written sampler as a reference corpus using loglikelihood.
- Conclusion: approval of the crossing paths between DH and CL as a necessity in the future of DH.

(3) Brooks, G. and T. McEnery (2020). Corpus linguistics. In *The Routledge Handbook of English Language and Digital Humanities*. edited by S. Adolphs and D. Knight. 378-404. Oxon, New York: Routledge.

CL and DH appear to be a good match: both are inextricably tied to digital technology, both use digital or digitalised data and both use computational tools for analysis.

2.2 ChatGPT and BARD

Q: What are key intersections and key roles that corpus linguistics has played in digital humanities?

A: ChatGPT¹⁾ vs. Bard's²⁾ answers

Corpus linguistics and digital humanities have a long and intertwined history, with numerous key intersections and roles. Here are some of the significant areas where corpus linguistics has impacted the digital humanities:

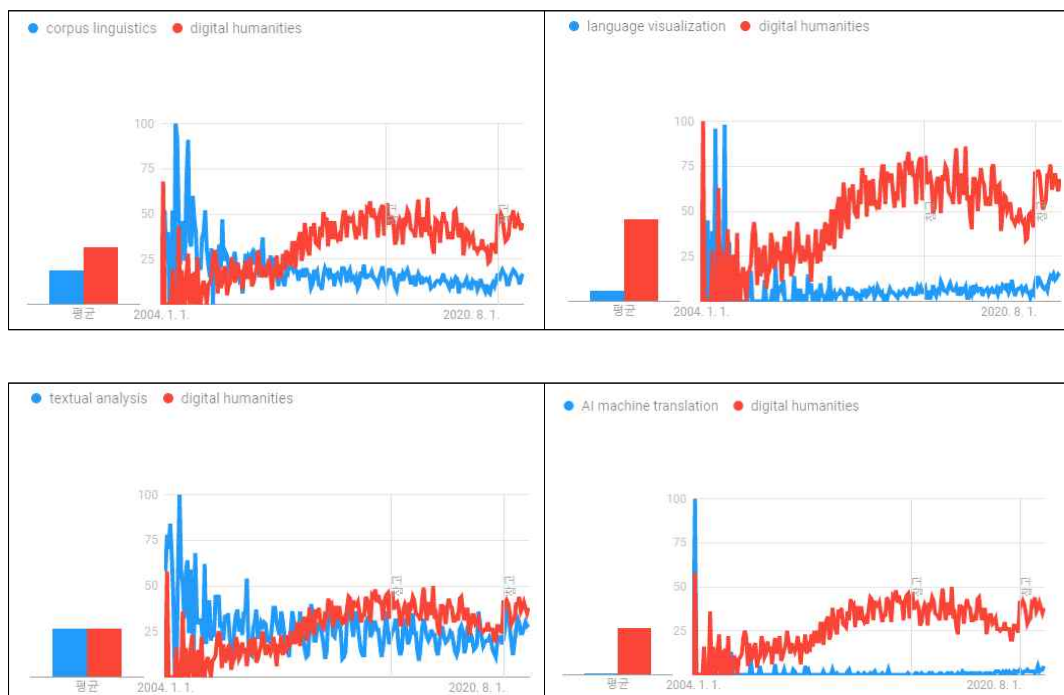
ChatGPT 4.0	Bard
Text Analysis and Visualization ³⁾	Intersection between CL and DH is the use of corpora to study historical language change and literary texts .
Natural Language Processing (NLP) ⁴⁾	CL and DH are used to study the use of other types of humanities texts , such as historical documents, legal texts, and religious texts.
Language Learning and Teaching ⁵⁾	CL provides a set of methods for analyzing large amounts of text data
Digital Editions and Scholarly Publishing	CL is used to develop new tools for digital humanities research. These tools can be used to analyze text data , create visualizations , and generate new insights into the structure, meaning, and history of humanities texts.
Digital Archiving and Preservation	
Cultural Analytics & Digital Lexicography	

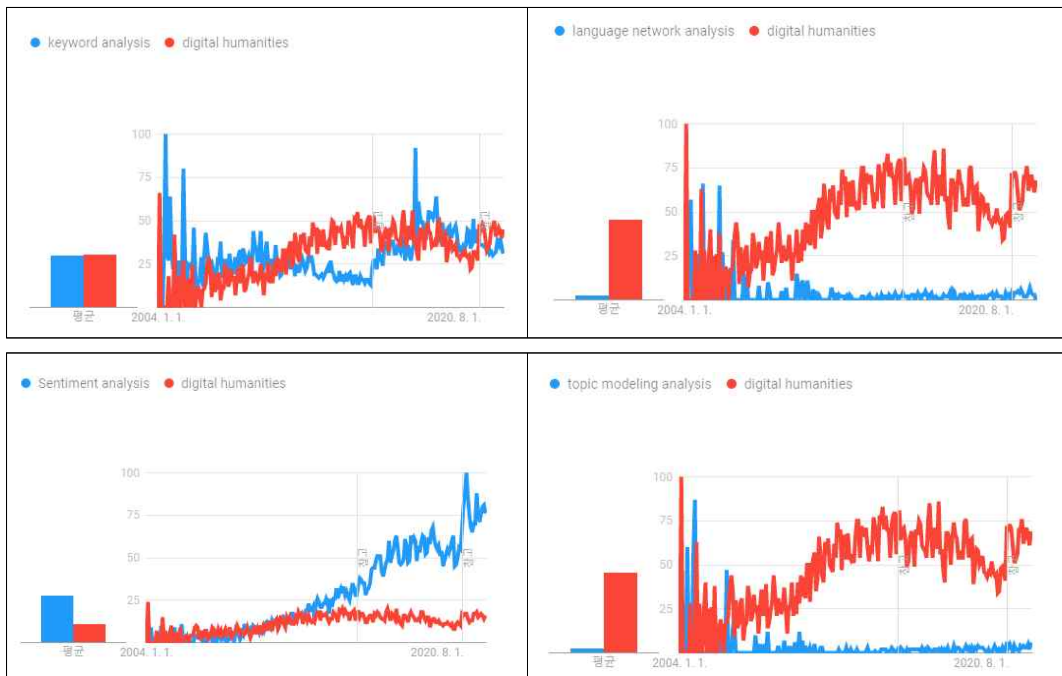
- 1) In all these ways and more, corpus linguistics plays a vital role in the digital humanities supporting the analysis, preservation, and communication of human culture in the digital age.
- 2) Corpus linguistics and digital humanities are two fields that are increasingly intersecting. Corpus linguistics is the study of large collections of text, called corpora, and the use of these corpora to investigate linguistic questions. Digital humanities is a broad field that uses digital

tools and methods to study humanities texts and artifacts

- 3) Corpus linguistics offers a suite of methods and techniques for analyzing large bodies of text. Techniques such as [frequency analysis](#), [collocation analysis](#), and [concordancing](#) have been applied to digital humanities projects to explore the thematic structure of large text collections, such as digitized books, newspapers, or social media posts. This kind of analysis also makes it possible to create [visualizations](#) that can help in understanding patterns, trends, and changes over time.
- 4) Corpus linguistics provides the foundation for NLP technologies. NLP technologies such as [sentiment analysis](#), [topic modeling](#), named entity recognition, and machine translation are used widely in the digital humanities to analyze text at scale
- 5) Corpus linguistics tools and methods are increasingly integrated into digital humanities projects that focus on language [learning and teaching](#). This includes the creation of learner corpora, which are collections of learner language that can be analyzed to better understand the learning process and develop more effective teaching materials and strategies.

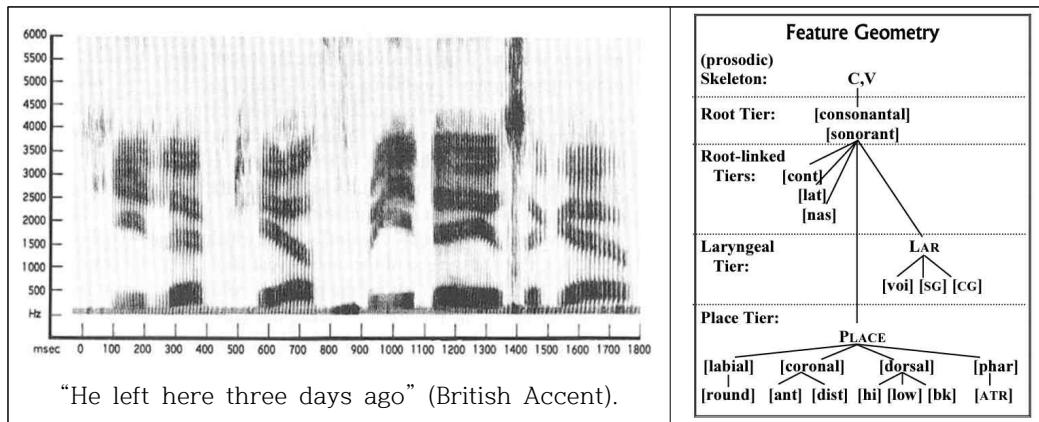
3. Trends in interest over time about the comparison of CL and DH (Google Trends from 2004 to 21 May, 2023)



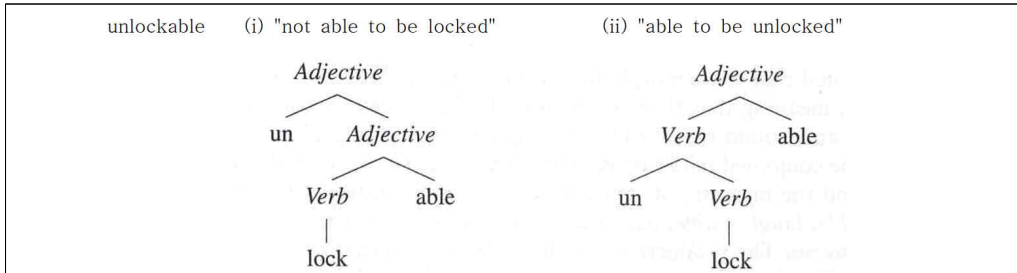


4. Visual representations in Linguistic theories

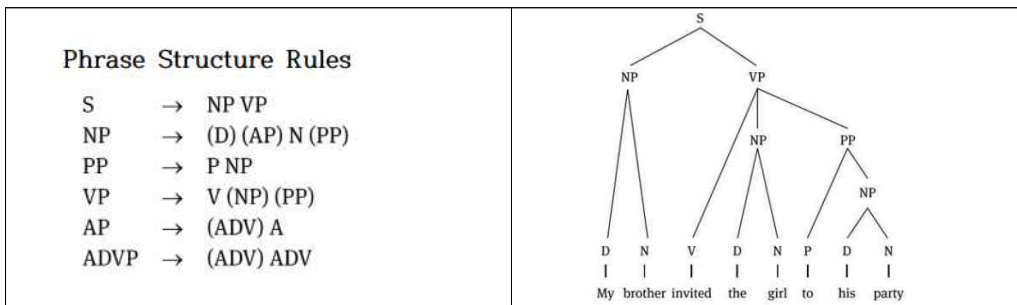
4.1 Sound representations (Spectrogram) and phonological features



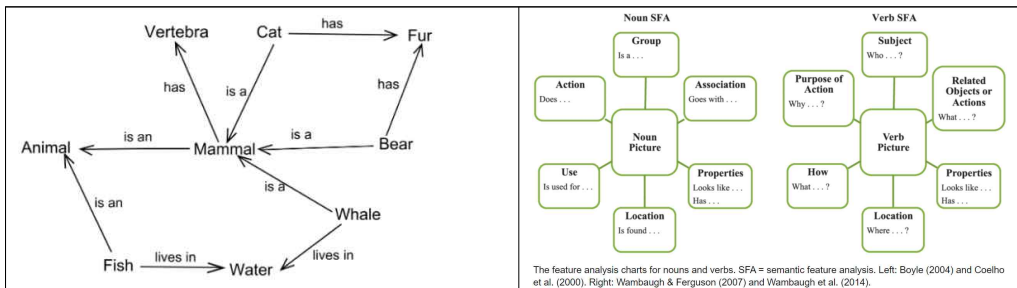
4.2 Internal structures of words



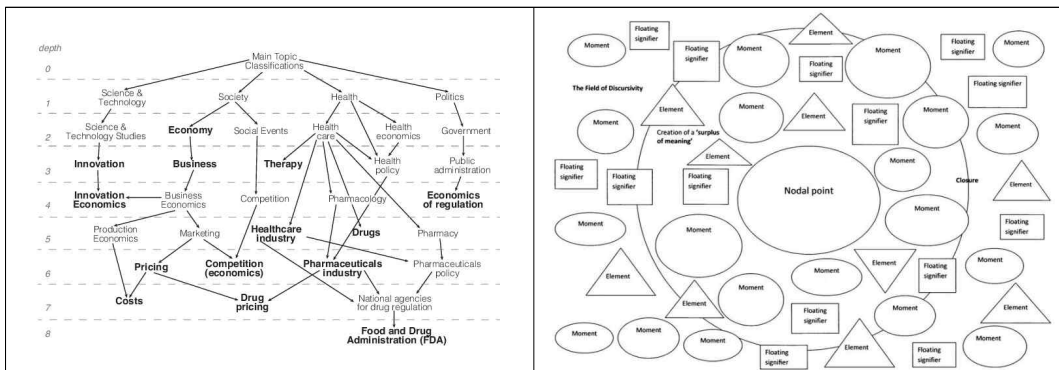
4.3 Syntax trees (phrase structure tree)



4.4 Semantic networks and semantic feature map



4.5 Discourse graphification and signs in discourse (structural visualization of discourse)



5. Case studies

5.1 Textual Analysis / AI machine translation

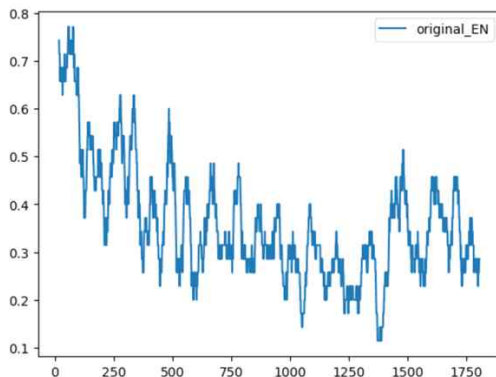
Liu, C., H. Park, and S. Jhang (2023) A Corpus-based Multi-lingual Comparison of AI-based Machine Translation: English, Korean, and Chinese. The on-going research paper that was presented at this conference today.

The aim of the present study is to investigate whether and to what extent the corpus linguistic technique, **type-token ratio (TTR)**, is able to **identify the quality of translation productions generated by different AI-based machine translation systems**. Specifically, this study will compare the discrepancy of Korean and Chinese translation outputs of one of James Joyce's short stories - 'Eveline,' generated by Google Translate (2006), DeepL (2017), and GPT-3.5 (2022), with reference to the corresponding original version written in English by visualizing their TTR.

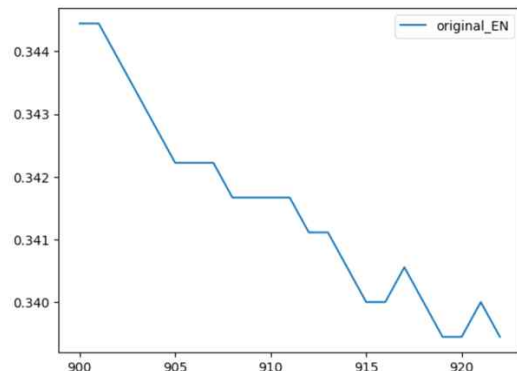
Two Research Questions:

1. Whether or not the translated Chinese and Korean TTR curves correlate with the three discourse boundaries claimed by literary critics?

Pilot study about the span



Original English *Eveline*, span=35

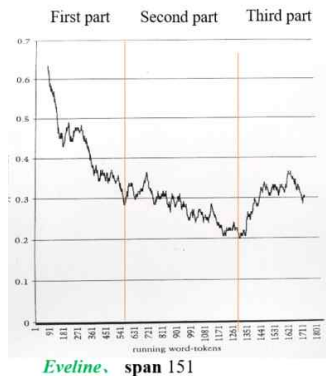


Original English *Eveline*, span=1800

- ⇒ The TTR calculated with a chosen span (English span 151 for 1,820 running words, Chinese span 131 for 1,807 running words, Korean span 231 for 2,794 running words) seems plausible to detect the significant boundaries in Chinese and Korean translations.

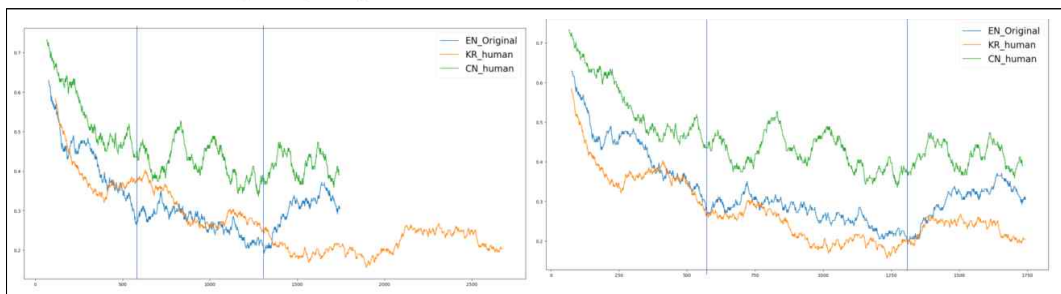
- Stubbs (2001) applied the method proposed by Youmans(1991) to examine the correspondence of the **text boundaries** in the **TTR chart** with the **literary critics** made by literature experts.

Literary critical interpretation vs. TTR (Eveline by James Joyce)

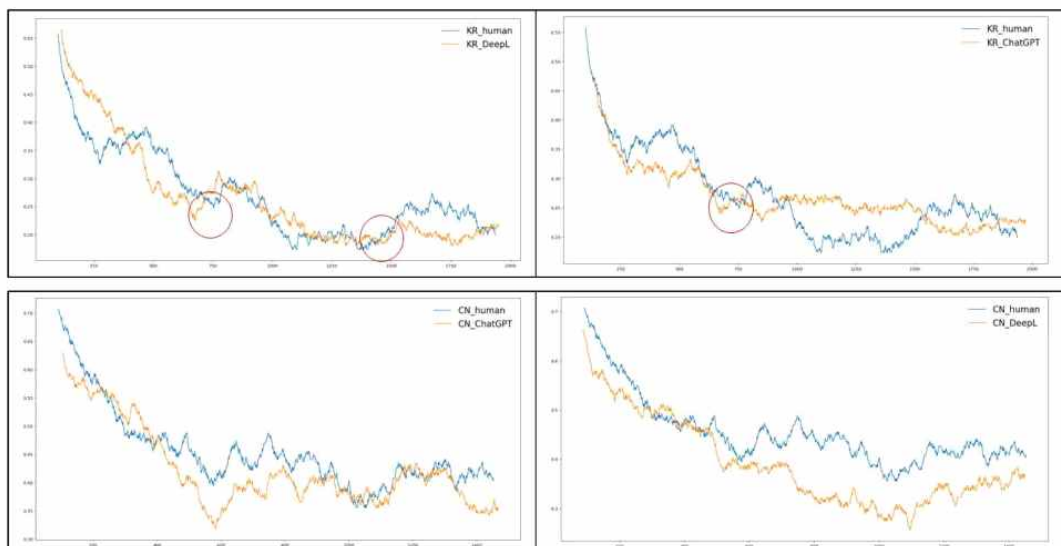


	First Part	Second Part	Third Part
Hart (1969)	Eveline thinks about the past and possible future	Brief interlude in which she reasserts her decision to choose life	Third part ends in her psychological failure
Stubbs (2001)	<ul style="list-style-type: none"> • Up-slope: High TTR means that a group of new vocabulary is introduced to make a new turn in the story. • “These three parts correspond astonishingly closely to the troughs and peaks identified by the program” • “The literary critic’s impressions are quite correct, but they can now be given an objective textual basis.” 		

The span is **an intuitive decision** with the aim to obtain interesting results which can reveal significant boundaries within text. If the span is too short, no detail and no clear patterns emerge. If it is too long, it will be too smooth (Stubbs, 2023 personal communication).



2. Do their TTR graphs show any discrepancies in the three AI-based machine translation systems such as Google Translate, DeepL, and GPT 3.5? How do we evaluate which is better or worse? ➡



Korean	Google	DeepL	GPT Playground
BLEU	0.2167 (3 rd)	0.2402 (1 st)	0.2213 (2 nd)
METEOR	0.3395 (2 nd)	0.3549 (1 st)	0.3176 (3 rd)
ROUGE	0.4347 (2 nd)	0.4438 (1 st)	0.3965 (3 rd)
Ranking	2nd	1 st	3rd

Evaluation of different metrics, Korean outputs

Chinese	Google	DeepL	GPT Playground
BLEU	0.1873 (1 st)	0.1790 (3 rd)	0.1795 (2 nd)
METEOR	0.3523 (2 nd)	0.3444 (3 rd)	0.3606 (1 st)
ROUGE	0.4971 (2 nd)	0.4767 (3 rd)	0.5156 (1 st)
Ranking	2nd	3rd	1 st

Evaluation of different metrics, Chinese outputs

Korean Ranking:
DeepL > Google > GPT

Chinese Ranking:
GPT > Google > DeepL

It seems that different MT models reveal different quality in terms of Chinese and Korean.

It seems that **the AI-based MTS are sensitive to particular languages.**

5.2 Collocation analysis and Keyword analysis

- ☐ Digital Humanities: Corpus and Literary Study
 - Corpus of literary works
 - e.g. Corpus of Shakespeare's draman
- ☐ Stylistic studies"
 - Compare the works of different writers"
 - Compare the literary works of different genres, for different readership



5.2.1 Collocation analysis

COLLOCATION is the relation between a word and individual word-forms which co-occur frequently with it (Lindquist, 2009: 57).

- **COLLIGATION** is the relation between a word and **grammatical** categories which co-occur frequently with it.
- **SEMANTIC PREFERENCE** is the relation between a word and **semantically** related words in a lexical field.
- **SEMANTIC PROSODY** is the **discourse** function of the word: it describes the speaker's communicative purpose.
 - Panic hasn't **set in** yet.
 - ... economic depression, lawlessness and authoritarianism **set in** ...
 - ... the postwar years when the cold war **set in**, ...
 - ... the psychological toll of unemployment has **set in**. (Lindquist, 2009: 58, from concordance lines for set in in COCA for the 2000s)

5.2.2 Keyword analysis

- Key literature that discusses features of Shakespeare's plays or key characters using keyword analysis: Busse (2002), Culpeper (2002), Scott and Tribble (2006), Murphy (2007), Archer, Culpeper and Rayson (2009), Verhaar (2016), Jhang, Se-Eun, Siqui Liu, and Jae-Hoon Kim (2018), etc.
- For example, Culpeper (2002) identifies concepts specifically relevant to six male and female characters in *Romeo and Juliet*, exploring keywords, key part-of-speech categories, key semantic categories and their role in text analysis.

Table 3.6 Top ten positive keywords for six characters in *Romeo and Juliet*.
frequency of occurrence given in brackets

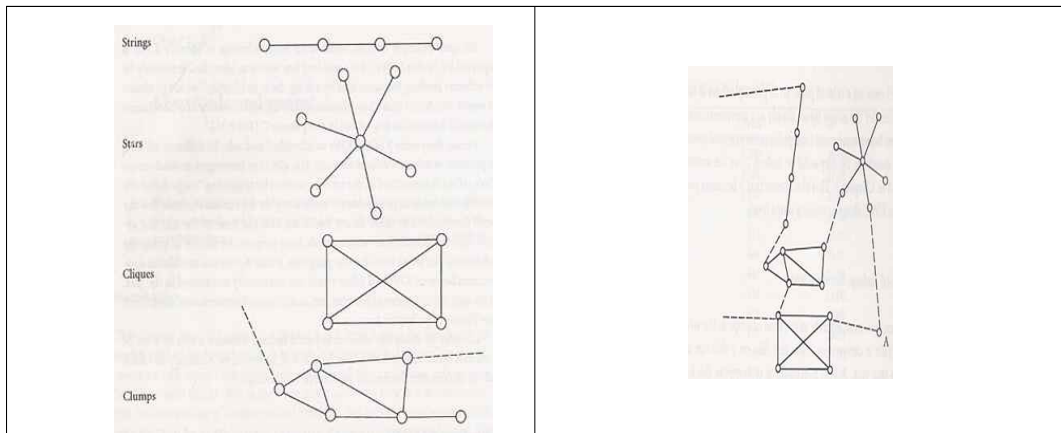
<i>Romeo</i>	<i>Juliet</i>	<i>Capulet</i>	<i>Nurse</i>	<i>Mercutio</i>	<i>Friar Laurence</i>
beauty (10)	if (31)	go (24)	day (22)	a (85)	thy (51)
blessed (5)	or (25)	wife (10)	he's (9)	hare (5)	from (23)
love (46)	sweet (16)	thank (5)	you (55)	very (11)	thysself (5)
eyes (14)	be (59)	ha (5)	quoth (5)	of (57)	Mantua (6)
more (26)	news (9)	you (49)	woeful (6)	he (20)	part (7)
mine (14)	my (92)	t (5)	God (12)	the (85)	heaven (10)
rich (7)	night (27)	Thursday (7)	warrant (7)	o'er (5)	forth (5)
dear (13)	I (138)	her (29)	Madam (10)		her (30)
yonder (5)	would (20)	child (7)	Lord (11)		alone (6)
farewell (11)	yet (18)	welcome (5)	Lady (16)		time (10)

Source: Based on Culpeper (2002: 19)

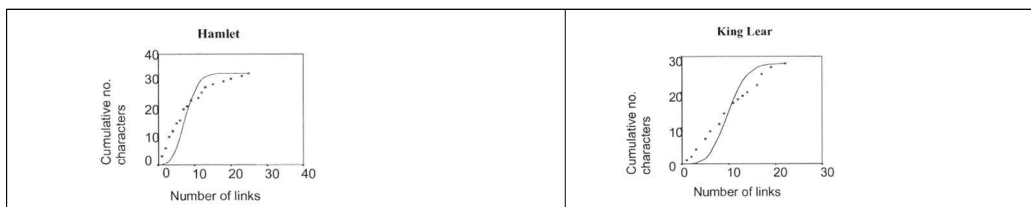
- Romeo' keywords match our picture of him as the lover of the play: *beauty*, *blessed*, *love* and *dear* belong in this sphere, as does the mention of body parts such as *eyes* and (further down the list, not reproduced here) *lips* and *hand*.
- Juliet' top keyword is *if*, and together with *yet* it suggests that Juliet is in a state of anxiety for much of the play.
- Capulet' top keyword is *go*, usually imperative and used in commands to members of his household, which is typical of his social role.
- The nurse is emotional (*woeful*, *God*) and either addresses or talks about people of higher status (*Madam*, *Lord*, *Lady*). She also likes to gossip about what other people have said (*quoth*).
- Mercutio' keywords indicate that he has a less interactive and more written-like style, with more noun phrases than the others (*a*, *of*, *the*).
- The fact that Friar Laurence has *thy* and *thysself* among the three top keywords leads Culpeper to a special investigation of the characters' use of personal pronouns, but we will not go into that here.

5.3 Language network analysis

- Scott and Tribble (2006: 75-76) shows four types of link between keywords nodes and a keyword linkage network adapted from Jones (1971: 56).



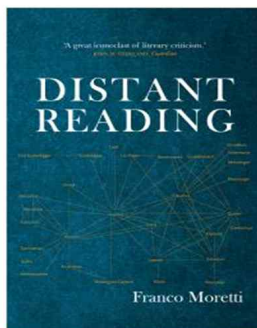
- See Jhang, S.E and S. Lee (2013) for literature review for language network and its visualization
- Network visualization as a digital humanities method for Shakespeare's plays
 - (1) Stiller, James, Daniel Nettle, & Robin I. M. Dunbar (2003) "The Small World of Shakespeare's Plays", *Human Nature*, Vol. 14, No. 4, 397-408.
- The mathematical analysis of networks as a basis for exploring the dramatic world of ten plays by William Shakespeare.
- The networks of interactions created by human beings to have “**small world properties**” because some networks have been shown to follow a scale-free distribution where the number of links per person follows a **power law**, with many individuals with a few links and a few individuals with many. (“the rich(small number)-get-richer and the poor(big number)-get-poorer”)



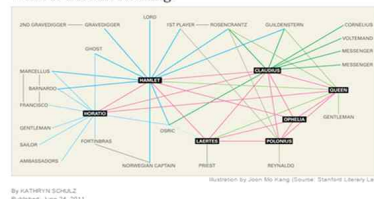
- Cumulative distribution of links per character (dots) as compared to a Poisson random model based on the mean number of links (line).
- (2) Stiller, J. and Hudson, M. (2005) Weak links and scene cliques within the small world of Shakespeare. *Journal of Cultural and Evolutionary Psychology*, 3(1), 57-73.
- Network analysis of 10 of Shakespeare's plays, comparing the structural features of the plays, the number of connections between each character, and the number of network connections a human observer would find.
- (3) Moretti, Franco (2011) *Network Theory, Plot Analysis*, Pamphlet 2, Stanford Literary Lab.

- Character-network as the so-called “small-world” property or “six degrees of separation”
- The network structure calculations were obtained by treating each speaking character as a vertex, and deeming two characters to be linked if there was at least one time slice of the play in which both were present (that is, if two characters spoke to each other or were in each other's presence, then they have a link).

(4) Moretti, Franco (2013) *Distant Reading*. London: Verso



What Is Distant Reading?



- How does a literary historian end up thinking in terms of z-scores, principal component analysis, and clustering coefficient?

(5) Wilhelm, Thomas, Manuel Burghardt, and Christian Wolff (2013) 'Visualising Hamlet'. 'To See or Not to See' - An Interactive Tool for the Visualization and Analysis of Shakespeare Plays. In ranken Wendelstorf, R., Lindinger, E. & Sieck J. (eds): *Kultur und Informatik - Visual Worlds & Interactive Spaces*, Berlin (pp. 75-185). Glückstadt: Verlag Werner Hülsbusch.

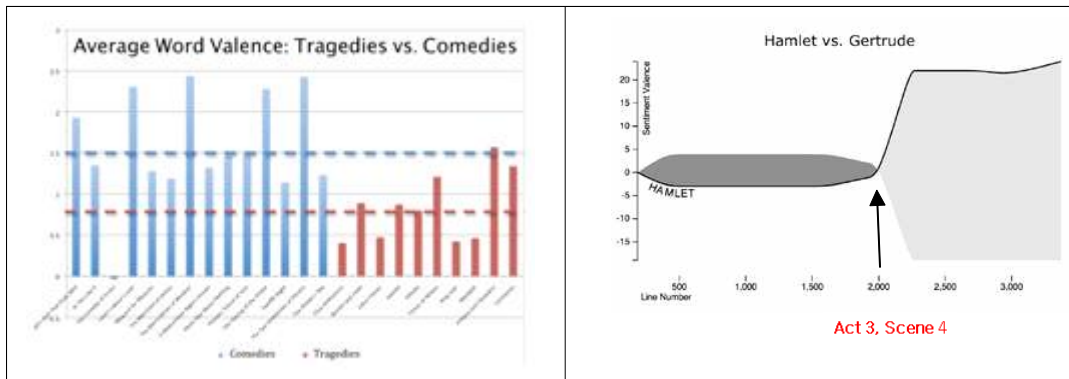
<http://www.thomaswilhelm.eu/shakespeare/output/hamlet.html>



(6) Nalisnick, Eric T. (2013) "Automatic Methods for Tracking Sentiment Dynamics in Plays", *Theses and Dissertations*, Paper 1569. (Master of Science in Computer Science and Engineering at Lehigh Univ.)

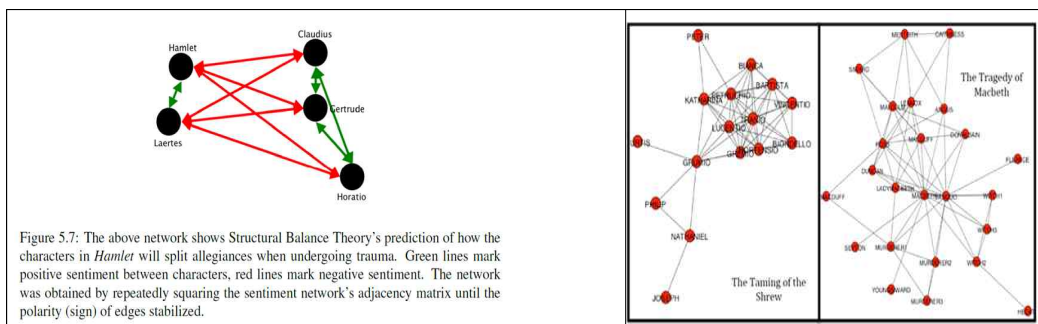
- A method for tracking the sentiment of the interpersonal relationships of literary characters to estimate the direction of sentiment flow, using AFINN.
- Sentiment can be tracked over the course of a work's plot to pinpoint **influential moments** as well as look for possibly hidden insights.

(7) Nalisnick, Eric T. & Henry S. Baird (2013a) "Extracting Sentiment Networks from Shakespeare's Plays", 12th International Conference on Document Analysis and Recognition, Washington, DC, 758-762.



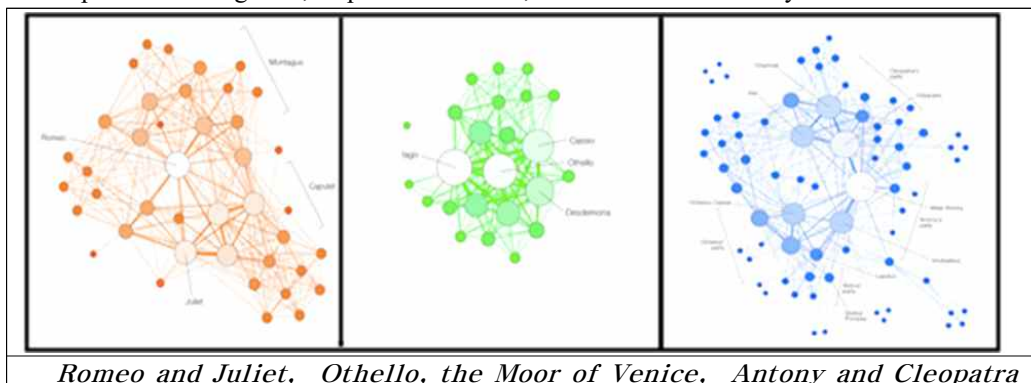
- (8) Nalisnick, Eric T.& Henry S. Baird (2013b) “Character-to-Character Sentiment Analysis in Shakespeare’s Plays, Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, 479–483.

- An automatic method for analyzing sentiment dynamics between characters in Shakespeare’s plays, using AFINN



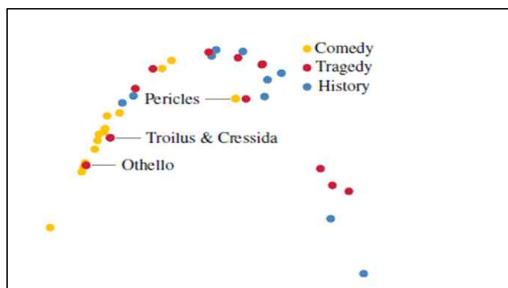
- (9) Grandjean, M. (2015) Network visualization: mapping Shakespeare's tragedies [Online] (Updated 23 Dec 2015) Available at: <http://www.martingrandjean.ch/network-visualization-shakespeare> [Accessed 10 April 2018]

He measured the density of the network structure between all characters in Shakespeare's 11 tragedies, expressed as a %, to see how close they are to each other.



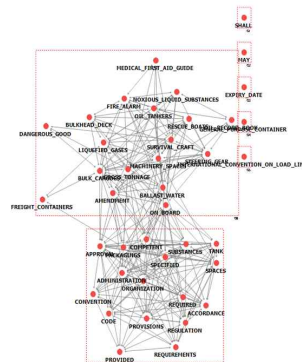
- (9) Rieck, Bastian & Heike Leitte (2016) "'Shall I compare thee to a network?'" Visualizing the Topological Structure of Shakespeare's Plays", Digital Humanities at IEEE Vis 2016, Baltimore, MD, USA.

Every **social network** describes co-occurrence relations between individual characters of a play; its weights may be used to indicate the importance of a connection between two characters, for instance.



- (10) Lee, S. (2016) Network Analysis of Maritime English Corpus with Multi-word Compounds: Keyword Networks and Collocation Network. Unpublished PhD dissertation, Korea Maritime and Ocean University.

He studied language networks using corpus linguistic data, adopting language network analysis and comparing keyword networks with collocation networks to identify **specific** or **general purpose** terms.

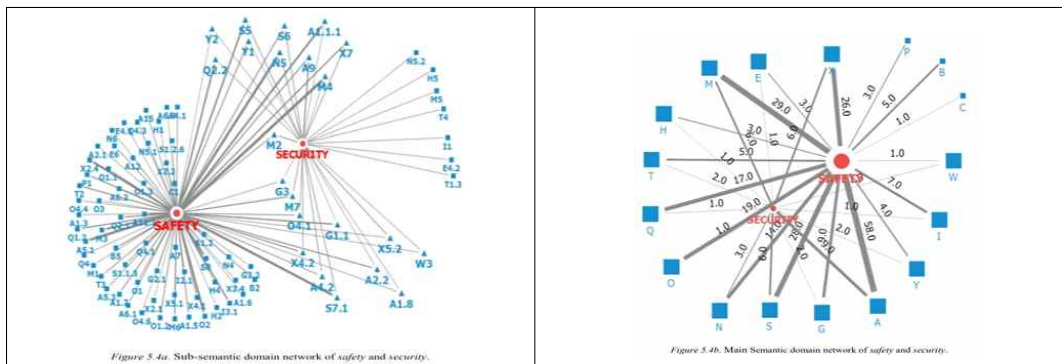


- (11) Lu, W. (2017) A Corpus-based Language Network Analysis of Near-synonyms in a Special Corpus. Unpublished PhD dissertation, Korea Maritime and Ocean University.

She examined 114 collocates of five groups of near-synonyms, i.e., shipvs. vessel, maritimevs. marine, oceanvs. sea, safetyvs. security, and harborvs. portdrawing data through WordSmith 6.0, tagging semantic domains in Wmatrix3.0 and conducting network analyses using NetMiner4.0, expecting to find a new and easy way to distinguish near-synonyms.

Table 5.3
Total number of collocates, sub-semantic domains, main semantic domains shared by each pair of near-synonym

Near-synonyms	Cosine similarity	Collocates	Sub-semantic domains	Main semantic domains
<i>vessel vs. ship</i>	0.4485	140	87	19
<i>maritime vs. marine</i>	0.2820	40	44	15
<i>sea vs. ocean</i>	0.2200	19	18	10
<i>safety vs. security</i>	0.1520	15	22	14
<i>harbor vs. port</i>	0.0736	6	16	8

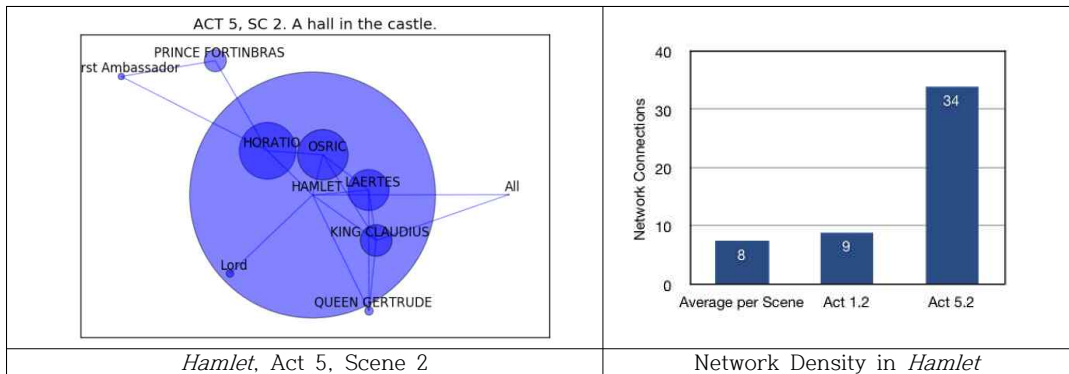


- (12) Lee, James and Jason Lee (2017) Shakespeare's Tragic Social Network; or Why All the World's a Stage, *Digital Humanities Quarterly*. Vol. 11 Issue 2, 201-232.

We propose that network visualization is a digital humanities method that can “explore” and “negotiate” the space between text and performance in the study of Shakespeare. The networks developed in this project use the language of Shakespearean plays to trace the relationships between characters in space, in effect, translating the literary text into a web of spatial relations, which are difficult to perceive solely in the act of reading. [...] This study represents the beginning of a digital method that aims to bridge text and performance in the study of Shakespeare by reading the dramatic text for the linguistic codes that organize the space of the stage

<p>THE PERSONS OF THE PLAY</p> <p>GHOST of Hamlet, the late King of Denmark</p> <p>KING CLAUDIUS, his brother</p> <p>QUEEN GERTRUDE of Denmark, widow of King Hamlet, now wife of Claudius</p> <p>Prince HAMLET, son of King Hamlet and Queen Gertrude</p> <p>POLONIUS, a lord</p> <p>LAERTES, son of Polonius</p> <p>OPHELIA, daughter of Polonius</p> <p>REYNOLDO, servant of Polonius</p> <p>HORATIO</p> <p>ROSENCRANTZ friends of Prince Hamlet</p> <p>GUILDENSTERN</p> <p>FRANCISCO</p> <p>BARNARDO soldiers</p> <p>MARCELLUS</p> <p>VALTEMAND</p> <p>CORNELIUS courtiers</p> <p>ORIC</p> <p>GENTLEMEN</p> <p>A SAILOR</p> <p>Two CLOWNS, a gravedigger and his companion</p> <p>A PRIEST</p> <p>FORTINBRAS, Prince of Norway</p> <p>A CAPTAIN in his army</p> <p>AMBASSADORS from England</p> <p>PLAYERS, who play the parts of the Prologue, Player King, Player Queen, and Lucianus, in 'The Mousetrap'</p> <p>Lords, messengers, attendants, guards, soldiers, followers of Laertes, sailors</p>	<p>ACT 1, SC 2. A room of state in the castle.</p>
Cast of Characters from the Oxford Shakespeare	<i>Hamlet</i> , Act 1, Scene 2.

To describe the **density of the network** in quantitative terms, Act 5.2 has 34 total network connections between characters, or 23.6% of total network links in the play as a whole, as opposed to the symmetrically organized scene of Act 1, Scene 2 above, which displays 9 network connections or 13% of the play's total network links.



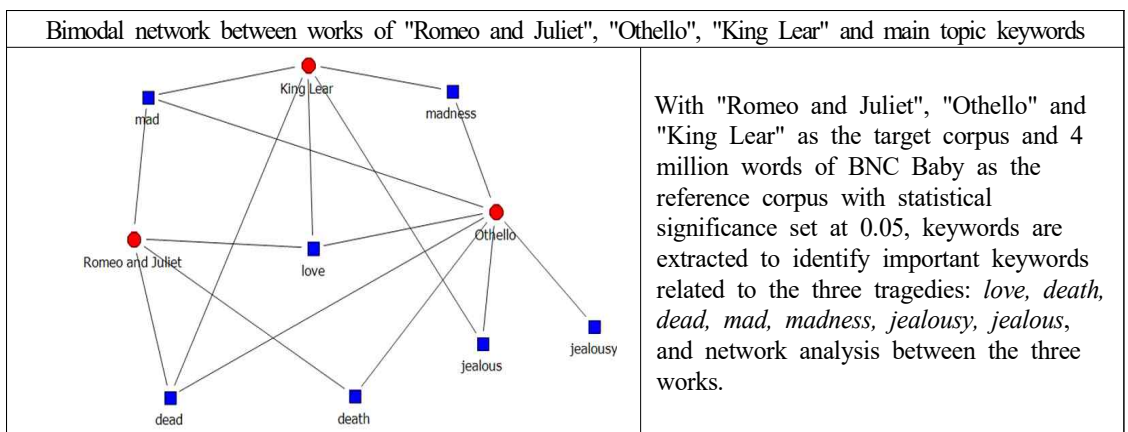
Act 5.2's 34 connections creates a network that is approximately 450% more dense than the play's average of 7.57 connections per scene.

- (13) **Proposal for the 2016 Selected General Joint Research Project** (Project No. 2016S1A5A2A03927611, 2016.11.01 ~ 2018.10.31, 2 years), "Keyword Network and Sentiment Analysis among Characters Using Dialogue Corpus of Shakespeare's Tragedies." National Research Foundation of Korea.

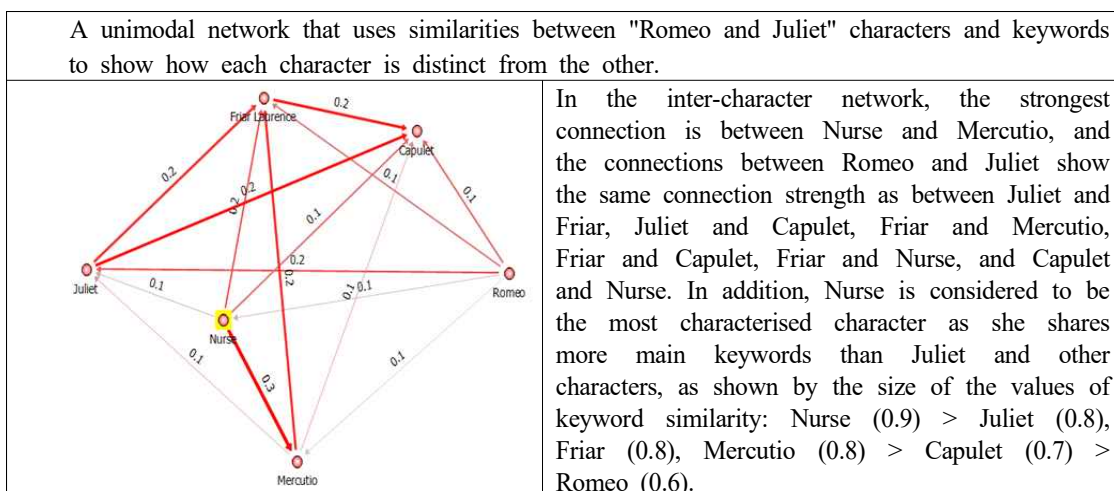
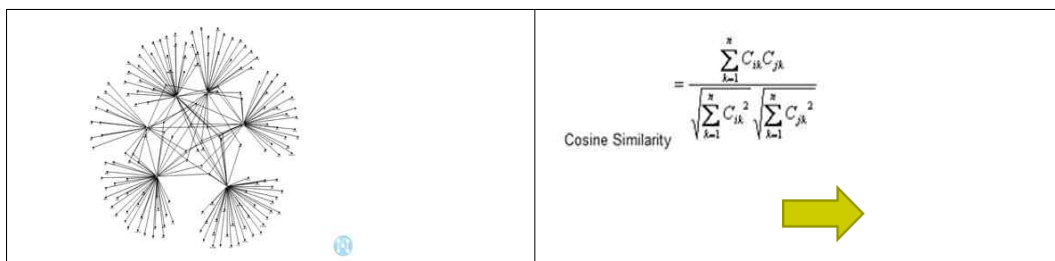
☐ Configuration of a dialogue network (using NetMiner4.0)

- Setting up one-mode between people, One Mode Network
- Tagging dialogue scenes
- A directed network from speaker to listener, with weights being the frequency of the conversation

☐ Visualisation of the four tragedies based on the level of connection in the conversation



☐ A character linkage network linked with whole keywords between six important characters in Romeo and Juliet



5.4 Sentiment analysis

5.4.1 Sentiment analysis of Shakespeare's tragedies

- (1) Jhang, S.E, Kyung-Eun Lee, Homin Park, Won-Moon Song, Hae-Ryong Jung, Soo-Sang Lee, and Jae-Hoon Kim (2019), Sentiment Analysis of Shakespeare's Tragedy Plays and Their Major Male and Female Characters through Dialogue Corpora. Journal of Language Sciences 26(1), 115-147

AFINN

- A sentiment dictionary with sentiment words manually labelled by Finn Arup Nielsen from 2009 to 2011, with each sentiment word assigned a value as an integer from -5 to 5.
- Latest version, AFINN-111, 2,477 words

VADER

- English sentiment analysis system adds [linguistic rules](#) to improve accuracy of dictionary-based sentiment analysis
- Create a sentiment dictionary using a 10-person collective intelligence-based method to verify accuracy and generality.
- The VADER (Gilbert and Hutto, 2014) sentiment dictionary contains 7,517 emotions, 3,345 positive emotions and 4,172 negative emotions, with each emotion assigned a mean of 0 as neutral and a real number between -4.0 and 4.0.
- Comparison of tragedy and comedy by AFINN vs. tragedy and comedy of love by BADER

- tragedy and comedy by AFINN

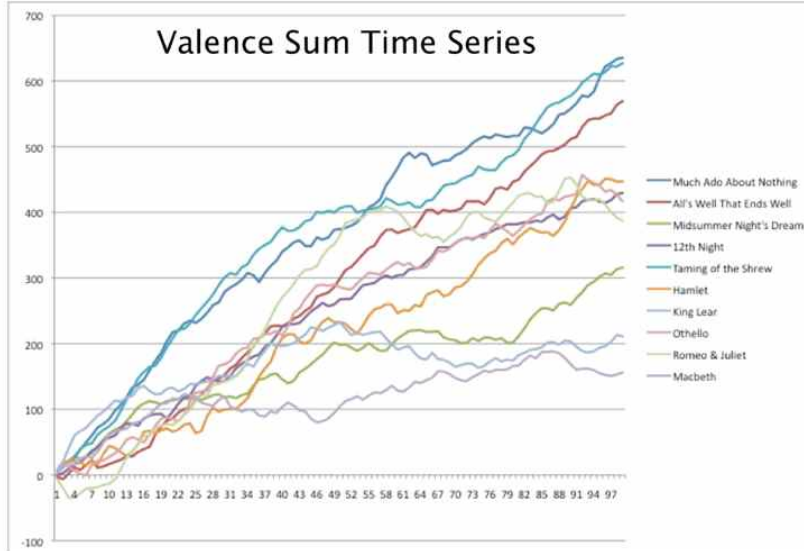
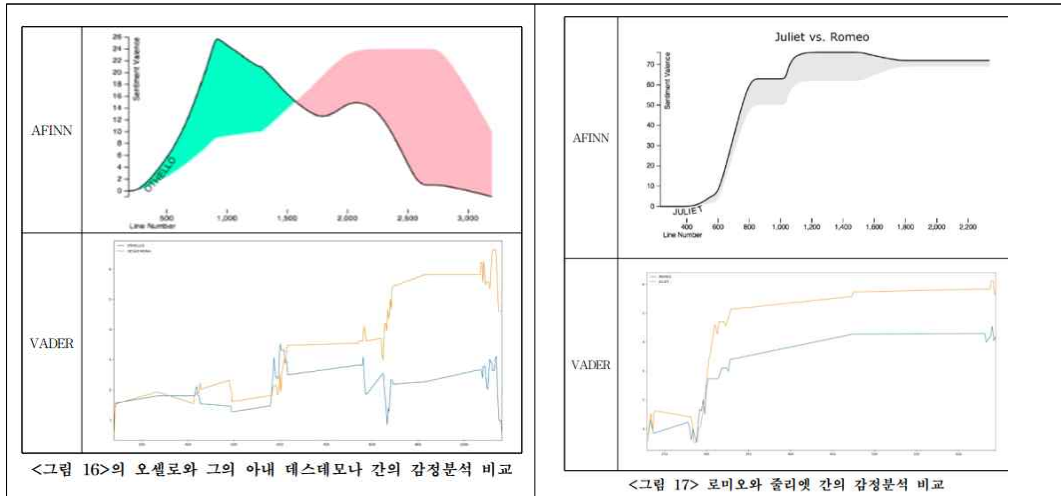


Figure 2.2: For each play, the valence value of each word was summed over time and was then normalized by breaking each play into percentiles. Tragedies tend toward the horizontal while comedies have a consistently positive slope.

- tragedy and comedy of love by BADER



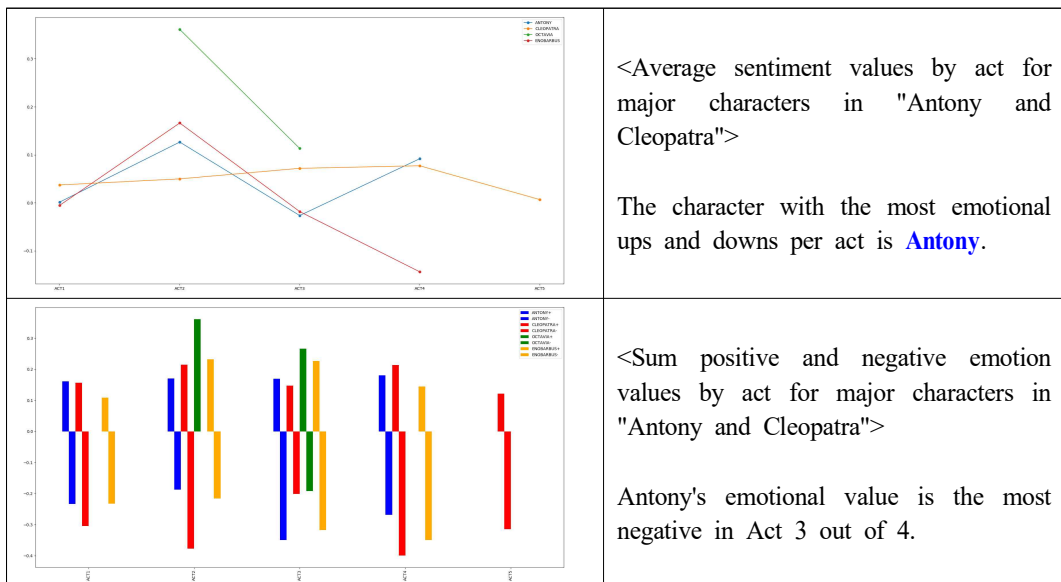
5.4.2 AFINN vs VADER sentiment analysis results by character in major tragedies



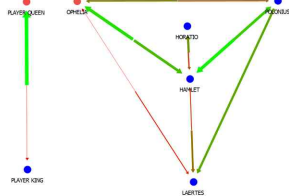
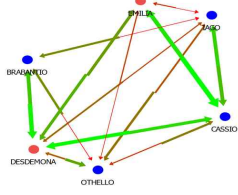
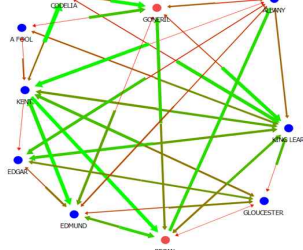
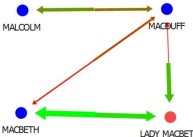
5.4.3 Sentiment analysis between works and between key characters in dialogue corpora

- (1) Analysis of the positive and negative sentiment of the entire film (overall, act-by-act, scene-by-scene) in line and bar graphs before performing sentiment analysis for each character.
- (2) Analysis of the emotions of individual characters within each work and the emotions conveyed between speaker and listener.

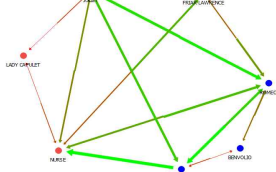
- Demonstration of how the main characters' emotions change from act to act
- Good observation of how the characters' emotions change as the play progresses.



Four Tragedies : The more narrow the **red** color is, the more negative the sentiment relation between characters is. The thicker **green** color is, the more positive the sentiment relation between characters is.

	<p>Hamlet spoke to four characters among six major characters in order of Ophelia > Polonius > Horatio at the positive tones but Laertes at the negative tones</p>
	<p>Othello spoke to just two major characters such as Cassio and Brabantio with positive tones and to others with negative tones.</p> <p>Desmeona's emotion toward Othello is much more positive than Othello's emotion toward Desmemona.</p>
	<p>King Lear spoke to seven characters except Edmund.</p> <p>King Lear spoke to six characters in order of (Gloucester > Goneril > Cordelia > Fool > Albany > Regan > Kent at the negative tones and to Edgar only with a positive tone.</p>
	<p>Macbeth spoke to just two characters such as Lady Macbeth and Macduff. He spoke to Lady Macbeth with a positive tone but to Macduff with a negative tone.</p>

Romeo and Juliet in Love Tragedies

	<p>In "Romeo and Juliet", "Romeo" has similar emotions to "Juliet".</p>
---	---

⇒ We tried to visualize network structures connecting major characters in Shakespeare's tragedy plays to easily figure out their detailed emotion relationship with regard to sentiment intensity (positive or negative), representing relation visualization. Adding expert opinion to the analysis of the degree of emotion between the protagonist and the main characters allows for new forms of analysis of existing works.

5.5 Topic modeling analysis

Jhang, Se-Eun, Soo-Sang Lee, Won-Moon Song, Hae Ryong Jung, Sung-Min Lee and Jae-Hoon Kim (2017) Topic Modeling Analysis in a Shakespeare Research Article English Abstract Corpus through Text Mining). *Journal of Language Sciences* 24(4), 55-82.

- Comparison of machine learning-based topic modeling keywords and corpus linguistics keywords
- Automatically extracting topics using topic modeling methodology for a corpus of English abstracts of domestic and foreign Shakespeare-related academic papers, and examining the top three topics that have been of particular interest and research through the changes in the total number of topics over the 37-year period from 1980 to 2016.
- Examination of all pattern types such as continuity, continuity, discontinuity, and one-offs of research topics up to the top 20 topics by year period, and the characteristics of downward, upward, and upward or downward changes in curvature.

- ✓ Analysed topics in four 10-year periods from 1980 onwards: 1980~1989 (101), 1990~1999 (208), 2000~2009 (474), and 2010~2016 (1,212).
- ✓ Topic modeling analysis with R
- ✓ Document Generation Criterion: 300 words in each document by using an R code which is provided by Thompson, Murakami, and Hunston (2016).
- ✓ Term Frequency-Inverse Document Frequency (TF-IDF) and Latent Dirichlet Allocation (LDA) algorithms were applied.
- ✓ Automatically generates an optimal 35 topics, each of which extracts a list of 30 keywords.

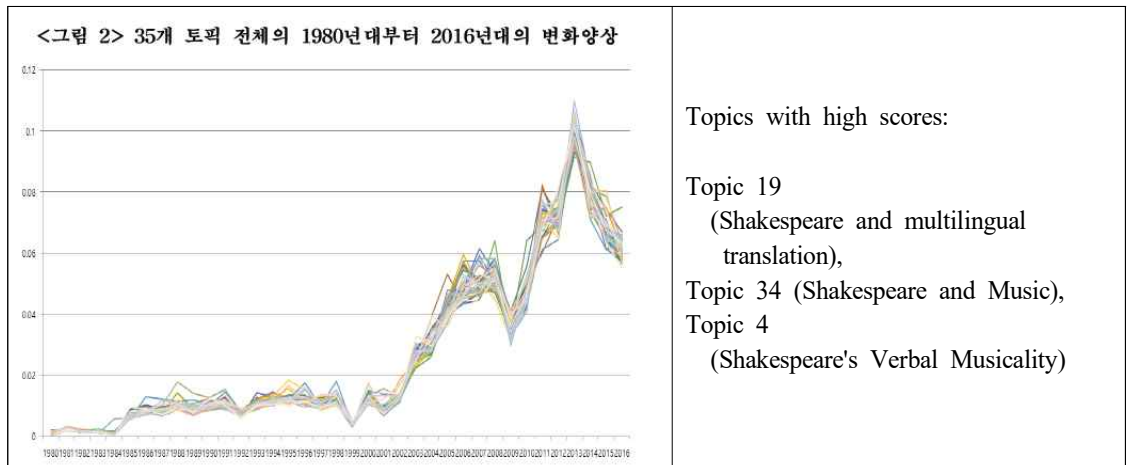
<30 comparative examples of absolute keywords and top 3 topic words for the period 2010-2016>

2010-2016	Topic 19	Topic 30	Topic 3
1	translation	japanese	fletcher
2	translator	marlowe	rhetoric
3	chinese	academic	rape
4	china	boy	falsehood
5	russian	humor	nationalism
6	arabic	japan	chaucer
7	citation	ego	commonplace
8	rendering	empire	oppositional
9	shakespearian	forgiveness	signature
10	persian	politeness	binary
11	arab	soliloquy	eloquence
12	researcher	length	lewis
13	biographical	manga	vocabulary
14	item	adolescent	opal
15	per	briefly	implicit
16	bengali	viola	stake
17	enlightenment	strange	usage
18	philippine	asia	censorship
19	phonological	brook	collaborator
20	team	feudal	don
21	troupe	blood	englishman
22	abroad	clock	formula
23	costume	closeness	pragmatics
24	irish	daydream	terrorism
25	kyle	guest	openness

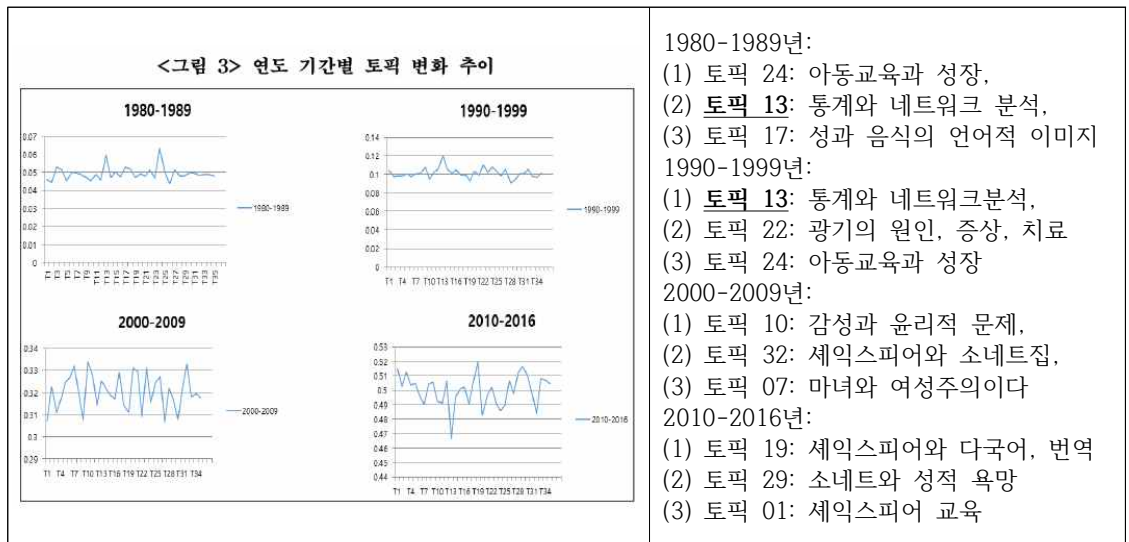
26	liang	publisher	strive
27	quarterly	thou	check
28	soviet	curse	expert
29	tam	listen	fair
30	terminology	maritime	lace

<Tagging and interpreting topics>

Topic	Labeling as Key Tag	Interpretation
01	셰익스피어 교육	셰익스피어 작품을 교육적으로 활용하는 방법과 교육자료 개발 방안
02	오셀로와 인종주의	『오셀로』(Othello)에 재현된 이종족에 대한 서구중심의 증오와 인종주의
03	셰익스피어의 극작 언어	셰익스피어와 협력집필의 관계에 있었던 존 플레처(John Fletcher)와 극작 언어
04	셰익스피어의 언어적 음악성	셰익스피어가 소네트와 극작품에 활용한 운문과 음악적 리듬
05	셰익스피어와 동성애적 심리분석	셰익스피어 작품에 대한 동성애적 심리분석
06	『태풍』과 후기식민주의 주제	『태풍』(The Tempest)에 표현된 후기식민주의(post colonialism) 주제와 내용
07	마녀와 여성주의	셰익스피어의 극화한 마녀와 마법에 관련된 여성주의적 주제
08	셰익스피어의 로마극과 영웅주의	『줄리어스 켄저』(Julius Caesar)에 극화된 남성적 우정과 로마적 영웅주의
09	셰익스피어와 경영 리더십	셰익스피어를 통해 경영의 기교와 리더십 모델을 발견하기
10	감성과 윤리적 문제	셰익스피어 작품에 나타난 감성과 윤리적 문제
11	민중주의적 정치성	귀족주의적 전통에 대한 조롱과 민중적 정치성
12	셰익스피어와 뉴미디어	셰익스피어의 작품을 새로운 형식의 미디어로 표현하는 문제
13	통계와 네트워크 분석	셰익스피어 저작물에 대한 통계적 분석 연구
14	식물의 약효와 생태학	셰익스피어 극에 등장하는 식물과 꽃이 가지는 약효와 생태학
15	질병과 지역	셰익스피어 극에 나타난 질병의 확산과 지역
16	상상적 은유적 표현에 대한 분석	셰익스피어의 상상과 은유적 표현 언어에 대한 분석
17	성과 음식의 언어적 이미지리	성과 음식에 대한 언어적 이미지리 연구
18	아시아 지역 풍토병	(셰익스피어와 관련성 모호) 아시아 지역 풍토병과 치료
19	셰익스피어와 다국어 번역	셰익스피어 작품의 다국어 번역 문제
20	언어학적 분석	셰익스피어 언어에 대한 언어학적 분석과 의미
21	태풍의 가부장주의와 충성의식	로맨스극 『태풍』(The Tempest)에 나타난 가부장 주의와 국가적 질서 의식
22	광기의 원인, 증상, 치료	광기와 그 치료에 대한 연구
23	글로벌극장과 공연환경	셰익스피어의 글로벌 극장과 공연환경 연구
24	아동교육과 성장	(셰익스피어와 관련성 모호) 아동의 훈육과 성장에 관한 연구
25	중세의 결혼과 자연의 섭리	중세의 결혼관과 자연의 섭리와 질서에 연구
26	대중매체와 대중 메시지	대중적 영상매체를 통해 재현된 셰익스피어 작품과 대중 메시지 연구
27	디지털 매체와 인간성	(셰익스피어와 관련성 모호) 디지털 시대의 인간성 문제와 차이에 대한 연구
28	여성에 대한 지배와 가부장주의	『말팔랑이 길들이기』에 담긴 가부장주의에 대한 정치적 종교적 해석과 여성에 대한 통치에 대한 문화적 개념연구
29	소네트와 성적 욕망	셰익스피어 소네트에 숨겨진 성과 동성애적 욕망
30	셰익스피어와 일본 만화	일본 만화로 표현된 셰익스피어와 크리스토퍼 말로우
31	사랑에 대한 시적 표현	(셰익스피어와 관련성 모호) 사랑, 구애, 늬움에 언어적 수사
32	셰익스피어 소네트집	셰익스피어 소네트집 출판에 대한 연구
33	세계 셰익스피어 페스티벌	셰익스피어 페스티벌에 대한 지역적 특성 연구
34	셰익스피어와 음악	셰익스피어 작품을 음악으로 재현하는 경향과 현상에 대한 연구
35	베니스 상인과 유대인	『베니스 상인』에서 재현한 사일록과 유대인에 대한 문제



□ Top three topics for each of the four 10-year time periods



- There was little overlap between these top three topics for each year period, but interestingly, only Topic 13 (Statistics and Network Analysis) was repeated in the first and second year periods, indicating that Topic 13 was the most covered topic over the 20-year period from 1980 to 1999.

6. Summary and future directions

6.1 Summary

Level	Corpus Linguistics	Digital Humanities
Words/Phrases	- Type-Token Ratio (TTR)	- Structures of Story - AI-based Machine Translation Systems
	- Lemma / Word-family	- Lexicography - Teaching language (vocabulary)
	- Collocation (Semantic prosody)	- Positive or negative emotion of speakers or writers
	- Keywords - Keyword-clusters - Lexical bundles - Key Clusters	- Aboutness of text - Styles of writers/speakers - Discourse analysis
	- Semantic annotation (Wmatrix)	- Literary analysis to identify whether certain semantic fields (like <i>war</i> , <i>love</i> , <i>religion</i> , etc.) are more prevalent in specific works or authors - Identity potential metaphorical expressions
	- Dialogue corpus - Sentiment dictionary	- Sentiment analysis of text and characters
Sentences	- Speech acts as pragmatic functions	- Literary analysis for the investigation of specific authorial styles in literature - Social media and online discourse analysis - Digital games studies
Paragraphs	- Topic modeling analysis	- Literary analysis to identify recurring themes in literary works - Trends in historical research, cultural studies, and other academic research
Register and Genre	- Language variety - Genre analysis	- Literary analysis to reveal how an author uses language differently when writing dialogue for different characters - Cultural studies to reveal patterns of language use specific to digital cultures

6.2 Future directions

The intersection of corpus linguistics and digital humanities continues to be a vibrant area of research, with many exciting potential future directions:

- **Interdisciplinary Collaborations:** One significant trend is the growing collaboration between linguists, computer scientists, historians, literature scholars, and other digital humanities

scholars. These collaborations can lead to new methods and research questions that combine the strengths of different disciplines.

- **Machine Learning and AI:** The application of machine learning and AI techniques to corpus analysis is another exciting area of development. This includes the use of deep learning for tasks such as [semantic role labeling](#), [sentiment analysis](#), [named entity recognition](#), and more.
- **Fine-Grained Semantic Analysis:** Tools like Wmatrix allow for semantic annotation of corpora, but there's still much work to be done in improving the accuracy and granularity of these tools. For instance, researchers are working on ways to automatically identify more subtle semantic phenomena, such as [metaphors](#) and idioms.
- **Multimodal Corpora**
- **Big Data Analysis**
- **Ethics and Bias in AI**
- **Longitudinal Studies**
- **More Diverse and Balanced Corpora (ChatGTP 4.0)**

—End—

Thank you very much for your attention.

Korean Middle School Third-Year English Textbook Reading Passage Difficulty Analysis: Focusing on Syntactic Complexity and Readability

고려대학교
윤인경, 유석훈

Table of Contents

1. Introduction
2. Methods
3. Result
4. Discussion

① Motivation

- It is appropriate to provide language materials that gradually introduce more complex vocabulary and sentences structures (Tomlinson, 2013).
- The analysis of learner proficiency level and alignment is an important evaluation criterion in textbook analysis (Mishan & Timmis, 2015).

② Claim

- There are differences in difficulty and comprehensibility based on reading passages by publishers, which may have statistical significance.

③ Significance

- Limited research has been conducted on the difficulty of reading passages in Korean middle school third-year English textbooks since the 2015 curriculum revision, indicating a need for further investigation and comparison with previous studies.

3

1. Do Korean middle school third-year English textbooks A, B, and C show significant differences in syntactic complexity?
2. Do Korean middle school third-year English textbooks A, B, and C show significant differences in readability?

① Korean English Textbook Data

- Three Korean middle school third-year English Textbooks based on the 2015 revised curriculum * (Three most widely used textbooks in Seongbuk-gu middle schools)
- The analyzed textbooks are identified using alphabetical symbols, as shown in Table 1

<Table 1> Lists of Korean middle school third-year English textbooks

Textbook	Publisher	Author	Number of Lessons
A	Chun-jae	Lee Jae-Young	8
B	Dong-A	Yoon Jeong-Mi	8
C	Neung-rile	Kim Seong-gon	8

5

2. Method

② Syntactic Complexity Analysis Tool

- The analysis of the extracted passages was conducted using the L2 Syntactic Complexity Analyzer (L2SCA), which calculated a total of 14 metrics across 5 categories to analyze the passages

<Table 2> L2SCA index

Category	Index	Formula
Structural complexity	Mean Length of Clause (MLC)	word ÷ clause
	Mean Length of Sentence (MLS)	word ÷ sentence
	Mean Length of T-unit (MLT)	word ÷ T-unit
Sentence complexity	Sentence Complexity Ratio (C/S)	clause ÷ sentence
Subordination	T-unit Complexity Ratio (C/T)	clause ÷ T-unit
	Complex T-unit Ratio (CT/T)	Complex T-unit ÷ T-unit
	Dependent Clause Ratio (DC/C)	Dependent clause ÷ clause
	Dependent Clause per T-unit (DC/T)	Dependent clause ÷ T-unit
Coordination	Coordinate Phrase per Clause (CP/C)	Coordinate phrase ÷ clause
	Coordinate Phrase per T-unit (CP/T)	Coordinate phrase ÷ T-unit
	Sentence Coordination Ratio (T/S)	T-unit ÷ sentence
Particular structure	Complex Nominal per Clause (CN/C)	Complex nominal ÷ clause
	Complex Nominal per T-unit (CN/T)	Complex nominal ÷ T-unit
	Verb Phrase per T-unit (VP/T)	Verb phrase ÷ T-unit

* L2SCA developed by Lu(2010, 2011)

2. Method

③ Readability Analysis Tools

- The readability of the textbook reading passages was analyzed using the number of sentences, words, words per sentence, vocabulary diversity, and readability indices such as Flesch-Kincaid Grade Level (FKGL) and Flesch Reading Ease (FRE).

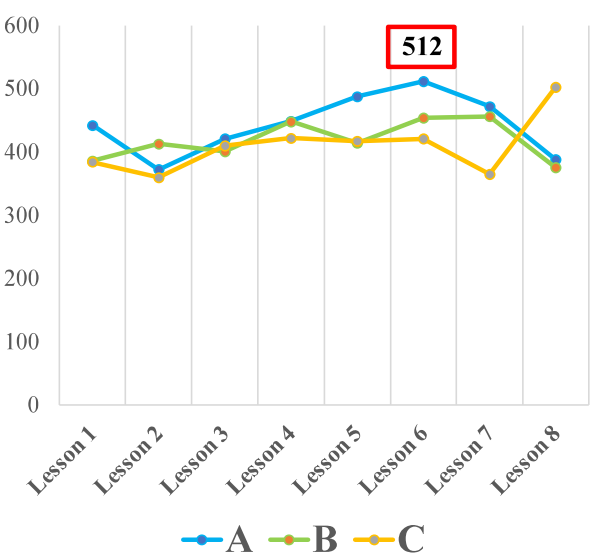
$$FKGL = (0.39 \times ASL) + (11.8 \times ASW) - 15.59$$

$$FRE = 206.835 - (1.015 \times ASL) - (84.6 \times ASW)$$

7

3. Result

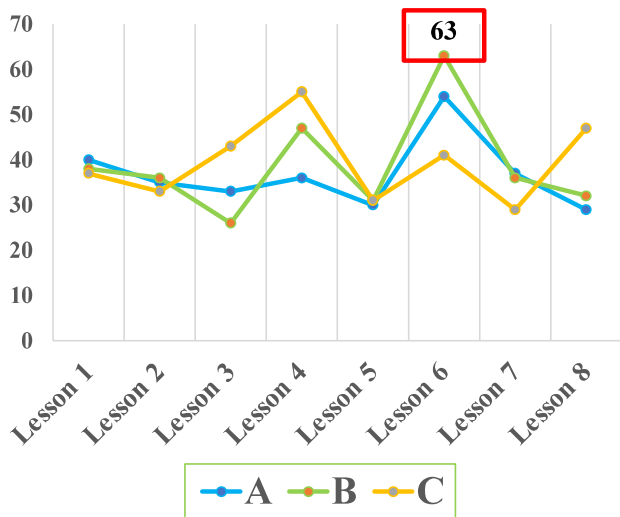
<Table 3> The number of words per unit in the Korean English textbooks



Textbook	A	B	C
Lesson 1	442	386	384
Lesson 2	372	413	360
Lesson 3	421	401	410
Lesson 4	449	448	422
Lesson 5	488	414	417
Lesson 6	512	454	421
Lesson 7	472	456	365
Lesson 8	388	375	503
Mean	443	418.37	410.25
SD	45.02	29.24	42.08

3. Result

<Table 4> The number of sentences per unit in the Korean English textbooks

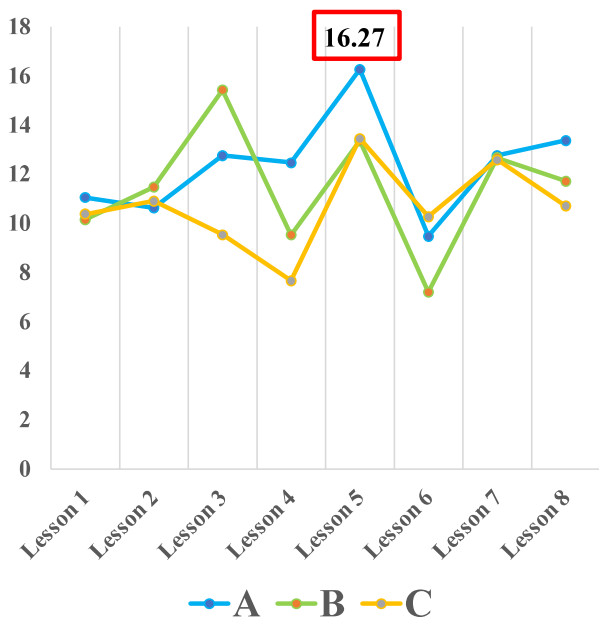


Textbook	A	B	C
Lesson 1	40	38	37
Lesson 2	35	36	33
Lesson 3	33	26	43
Lesson 4	36	47	55
Lesson 5	30	31	31
Lesson 6	54	63	41
Lesson 7	37	36	29
Lesson 8	29	32	47
Mean	36.75	38.62	39.5
SD	7.34	10.84	8.23

9

3. Result

<Table 5> The average sentence length per unit in the Korean English textbooks



Textbook	A	B	C
Lesson 1	11.05	10.16	10.38
Lesson 2	10.63	11.47	10.91
Lesson 3	12.76	15.42	9.53
Lesson 4	12.47	9.53	7.67
Lesson 5	16.27	13.35	13.45
Lesson 6	9.48	7.21	10.27
Lesson 7	12.76	12.67	12.59
Lesson 8	13.38	11.72	10.70
Mean	12.35	11.44	10.69
SD	1.92	2.35	1.66

3. Result

<Table 6 > Comparison of mean text lengths based on clauses, sentences, and T-units

지표	자료	n	평균/표준편차	F	p	사후분석(Scheffe)	
						집단	평균차이
MLC	A	8	8.22/1.34	0.28	.000***	a-b	0.11
	B	8	7.99/1.29			b-c	0.16
	C	8	7.71/1.22			a-c	0.55
MLS	A	8	12.35/1.92	1.21	.000***	a-b	0.72
	B	8	11.44/2.35			b-c	0.5
	C	8	10.69/1.66			a-c	2.42
MLT	A	8	11.12/1.51	0.52	.000***	a-b	0.42
	B	8	10.51/1.93			b-c	0.13
	C	8	10.18/1.82			a-c	1.01

* MLC (Mean Length of Clause); MLS (Mean Length of Sentence); MLT (Mean Length of T-unit)

11

3. Result

<Table 7> Comparison of sentence complexity

지표	자료	n	평균/표준편차	F	p	사후분석(Scheffe)	
						집단	평균차이
C/S	A	8	1.51/0.19	1.18	.000***	a-b	1.18
	B	8	1.42/0.13			b-c	0.16
	C	8	1.39/0.13			a-c	2.21

* C/S (Clause per Sentence)

3. Result

<Table 8> Comparison of subordinations based on the complexity of T-units, complex T-units, dependent clauses, and dependent clauses per T-unit

지표	자료	n	평균/표준편차	F	p	사후분석(Scheffe)	
						집단	평균차이
C/T	A	8	1.37/0.18	0.29	.000***	a-b	0.49
	B	8	1.32/0.11			b-c	0.01
	C	8	1.32/0.13			a-c	0.37
CT/T	A	8	0.35/0.13	0.38	.000***	a-b	0.66
	B	8	0.30/0.10			b-c	0.02
	C	8	0.31/0.09			a-c	0.47
DC/C	A	8	0.26/0.07	0.07	.000***	a-b	0.01
	B	8	0.26/0.09			b-c	0.70
	C	8	0.27/0.07			a-c	0.14
DC/T	A	8	0.36/0.15	0.03	.000***	a-b	0.03
	B	8	0.35/0.13			b-c	0.05
	C	8	0.37/0.12			a-c	0.02

* C/T (T-unit Complexity Ratio); CT/T (Complex T-unit Ratio); DC/C (Dependent Clause Ratio); DC/T (Dependent Clause per T-unit)

13

3. Result

<Table 9> Comparison of coordinations based on phrase and sentence ratios

지표	자료	n	평균/표준편차	F	p	사후분석(Scheffe)	
						집단	평균차이
CP/C	A	8	0.13/0.05	0.87	.000***	a-b	0.35
	B	8	0.15/0.07			b-c	0.52
	C	8	0.19/0.11			a-c	1.72
CP/T	A	8	0.17/0.05	0.90	.000***	a-b	0.30
	B	8	0.20/0.09			b-c	0.63
	C	8	0.24/0.15			a-c	1.79
T/S	A	8	1.11/0.05	1.07	.000***	a-b	0.37
	B	8	1.09/0.10			b-c	0.72
	C	8	1.06/0.05			a-c	2.12

* CP/C (Coordinate Phrase per Clause); CP/T (Coordinate Phrase per T-unit); T/S (Sentence Coordination Ratio)

3. Result

<Table 10> Comparison of particular structures based on clauses and T-units

지표	자료	n	평균/표준편차	F	p	사후분석(Scheffe)	
						집단	평균차이
CN/C	A	8	0.71/0.18	0.42	.000***	a-b	0.06
	B	8	0.74/0.23			b-c	0.79
	C	8	0.65/0.16			a-c	0.40
CN/T	A	8	0.96/0.20	0.45	.000***	a-b	0.03
	B	8	0.98/0.34			b-c	0.8
	C	8	0.85/0.24			a-c	0.53
VP/T	A	8	1.67/0.28	0.69	.000***	a-b	0.44
	B	8	1.59/0.17			b-c	1.37
	C	8	1.73/0.20			a-c	0.25

* CN/C (Complex Nominal per Clause); CN/T (Complex Nominal per T-unit); VP/T (Verb Phrase per T-unit)

15

3. Result

<Table 11> Comparison of readability based on FKGL and FRE scales

지표	자료	n	평균/표준편차	F	p	사후분석(Scheffe)	
						집단	평균차이
FKGL	A	8	5.70/1.52	0.22	.000***	a-b	0.44
	B	8	5.16/1.83			b-c	1.37
	C	8	5.56/1.31			a-c	0.25
FRE	A	8	75.29/8.80	0.33	.000***	a-b	0.21
	B	8	77.41/9.28			b-c	0.66
	C	8	73.64/7.88			a-c	0.13

* FKGL (Flesch-Kincaid Grade Level); FRE (Flesch Reading Ease)

3. Result

1. Do Korean middle school third-year English textbooks A, B, and C show significant differences in syntactic complexity?
 - Textbook A exhibited the highest average clause count in sentences with a mean of 1.51, whereas Textbooks B and C which had averages of 1.42 and 1.39.
 - The analysis of syntactic complexity revealed that the greatest disparity in difficulty was observed between Textbooks B and C.

17

3. Result

2. Do Korean middle school third-year English textbooks A, B, and C show significant differences in readability?
 - Based on the revised 2015 curriculum, the average FKGL readability score of the three selected textbooks was 5.47, indicating a difficulty level comparable to that of students in grades 5 to 6 in the United States (Kincaid, 1975).
 - In a previous study conducted by Bae (2019), which analyzed a total of eight textbooks for the 1st and 2nd grades of middle school, the FKGL readability scores were found to be 2.7 and 3.7, respectively.
 - The overall findings indicate that the difficulty level of middle school textbooks increases by approximately twice as much from the first and second years to the second and third years.

4. Discussion

Complex Nominal Structures

- Textbook B: Highest complexity (CN/C and CN/T) - Lesson 3 (1.13)

Verb Phrases per T-Unit

- Textbook C: Highest complexity (VP/T) - Lesson 5 (2.16)

Syntactic Complexity and Readability

- Textbook A: Highest complexity (C/T) - Lesson 1 (1.8)
- Textbook A: Highest mean FRE score, suggesting consistent readability (72.3)
- Textbook B: Lowest FKGL score – Lesson 6 (1.4)

Implications

- Consider linguistic features when selecting or designing textbooks.
- Balance challenge and support based on proficiency levels and learning objectives.
- Scaffold and differentiate instruction accordingly.
- Use materials with varying readability levels to meet diverse needs.

17

Reference

Brown, H. D., & Lee, H. (2014). *Teaching by principles* (4th ed.). New York: Longman.

Brown, J. D. (1998). An EFL readability index. *JALT Journal*, 20(20), 171-200.

Fry, E. (1977). Fry's readability graph: Clarification, validity, and extension to level 17. *Journal of Reading*, 20, 242-252.

Greenfield, J. (2004). Readability formulas for EFL. *JALT Journal*, 26(1), 5-24.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (Automated readability index, Fog count and Flesch reading ease formula) for navy enlisted personnel (No. RBR-8-75). Millington, TN; Institute for Simulation and Training.

McDonough, J., Shaw, C., & Masuhara, H. (2013). *Materials and methods in ELT: A teacher's guide* (3rd ed.). Oxford: Wiley-Blackwell.

McGrath, I. (2016). *Materials evaluation and design for language teaching* (2nd ed.). Edinburgh, England: Edinburgh University Press.

Tomlinson, B. (Ed.) (2013). *Applied linguistics and materials development*. London: Bloomsbury.

김성곤, 서성기, 이석영, 최동석, 강용구, 김성애, 최인철, 양빈나, 김지연, 이설미(2020). *Middle school English 3*. 서울: 엔이능률.

윤정미, 이희경, 송형호, 염미선, 진성인, 장성욱, Sundeen Glenn Paul(2020). *Middle school English 3*. 서울: 동아출판.

이재영, 안병규, 오준일, 배태일, 김순천, 신수진, 박리원(2020). *Middle school English 3*. 서울: 천재교육.

Anti-Clitics and Syntactic Derivation in Korean: Focussing on Lexico-Morpho-Syntax of *kes*, *nom* and *ca*

Korean Association for Corpus Linguistics (KACL) ·
Korean Association for the Study of English Language and Linguistics (KASELL)
Joint Summer Conference on “현대 언어학의 현황과 과제: 코퍼스, 빅데이터, 인공지능”
Sat., June 3, 2023, Korea University



1. Objectives, Issues, and Basic Ideas



1. Objectives, Issues, and Basic Ideas [1/6]

- ① (1)은 (2b)와 (2c)로 해석되지만, 주문장 술어의 선택제약을 만족시키지 못하기 때문에 (2a)로는 해석되지 않는다.

(1) 철수는 [[영희가 마늘을 다진] 것]을 보관하였다.

(2) a. # 철수는 [마늘을 다진] 영희를 보관하였다.

b. 철수는 [영희가 다진] 마늘을 보관하였다.

(대상)

c. 철수는 [영희가 마늘을 다진] 도구를 보관하였다.

(도구)

- ② 내핵관계절 가설(= 내부에 명시적으로 실현되어 있는 명사구가 피봇 명사로 기능하는 명사절)의 검증 현상

- Jhang(1994), Park(1994), Jung(1995), Y.-B. Kim(1996; 2002), J.-B. Kim(1999), Chung & Kim(2003), M.-J. Kim(2004a; 2004b; 2007), Cha(2005), Kim, Sells & Yang(2006), J. Lee(2006), Song(2021), K.-S. Kim(2022), Ryu(2022), Ryu(2023), etc.
- (1)이 (2b)로 해석되는 현상은 내핵관계절 가설로 분석할 수 있다고 해도
- 도구 명사구는 명시적으로 실현되지 않았기에 내핵관계절 가설로도 (1)이 (2c)로 해석되는 현상은 분석할 수 없다.

1. Objectives, Issues, and Basic Ideas [2/6]

- ③ '것'이 도구로 해석되는 경우 수식절 안에 도구 명사가 실현되지 않으면 문법적이나 오히려 모두 실현되면 비문법적이다.

(3) a. 철수는 [[영희가 마늘을 다진] 것]으로 돼지고기도 다졌다. (도구)

b. *철수는 [[영희가 무쇠절구통으로 마늘을 다진] 것]으로 돼지고기도 다졌다. (도구)

- ④ '것'이 대상으로 해석되는 경우 수식절 안에 대상 명사가 실현되지 않으면 문법적이나 오히려 모두 실현되면 비문법적이다.

(4) a. 철수는 [[영희가 무쇠절구통으로 다진] 것]을 넣고 김치찌게를 끓였다. (대상)

b. *철수는 [[영희가 무쇠절구통으로 마늘을 다진] 것]을 넣고 김치찌게를 끓였다. (대상)

- ⑤ 내포절 술어와 관련된 어떠한 논항도 실현되지 않은 경우에 도구 해석과 대상 해석은 가능해서 (5)는 중의적이다.

(5) 철수는 [[다진] 것]을 보관하였다.

a. 대상: 마늘, 쇠고기, 돼지고기 등

b. 도구: 마늘다지기, 무쇠절구, 핸드블렌더, 갈릭프레스, 믹서, 식칼 등

1. Objectives, Issues, and Basic Ideas [3/6]

⑥ Similarities and Differences between 것, 놈, and 자

- (6) a. 철수는 [[영희가 마늘을 다진] {것을/놈을/*자를}] 찾았다.
 b. 철수는 [[마늘을 다진] {것을/놈을/자를}] 찾았다.
 c. 철수는 [[다진] {것을/놈을/자를}] 찾았다.
- (7) a. 다진 {
 것 (= 행위, = 대상, = 도구, ≠ 행위자)
 놈 (≠ 행위, = 대상, = 도구, = 행위자)
 자 (≠ 행위, ≠ 대상, ≠ 도구, = 행위자)}
- b. 마늘을 다진 {
 것 (= 행위, = 대상, = 도구, ≠ 행위자)
 놈 (≠ 행위, ≠ 대상, ≠ 도구, = 행위자)
 자 (≠ 행위, ≠ 대상, ≠ 도구, = 행위자)}
- c. 영희가 마늘을 다진 {
 것 (= 행위, ≠ 대상, = 도구, ≠ 행위자)
 놈 (≠ 행위, ≠ 대상, = 도구, ≠ 행위자)
 *자 (≠ 행위, ≠ 대상, ≠ 도구, ≠ 행위자)}

1. Objectives, Issues, and Basic Ideas [4/6]

? 문제 제기

- '것'과 선행절 결합체의 해석에 논항의 명시적 실현 조건은 너무 강하거나 설명에 제한적이다.
- 논항의 명시적 실현 조건의 완화 혹은 폐기는 어떤 논항도 실현되지 않은 문장 (5)와 같은 명사 수식절을 모두 내핵관계절로 간주해야 함을 의미하기에 진퇴양난

i 기본 아이디어

- '것'과 선행절 결합체의 해석에 논항의 실현 여부는 상관이 없고, 기본적으로 선행절 술어의 논항구조와 관련되어 있다.
- '것'은 이들 중 특정 논항을 실현시켜 명사로 파생시키는 기능을 담당한다 (cf. *destruction*, *examinee*, *mixer*).
- 이 명사 파생 기능은 파생 접사(affix)로서도 아니고 음운론적 자립 단어(word)로서도 아니고 반접어(anti-clitic)으로서 수행한다.
- 반접어(anti-clitic): 음운론적으로 독립적인 단어이나 역할은 접사와 같이 명사 파생 역할

i 분석의 형식적 이론 토대

- Recent development of the lexicon theory in Head-driven Phrase Structure Grammar (HPSG)
- Pollard & Sag(1994), Sag, Wasow & Bender(2003), Müller, Abeillé, Borsley & Koenig (eds.) (2021). etc.

1. Objectives, Issues, and Basic Ideas [5/6]

✓ Goals & Proposal

- ① Introduction of the notion of “anti-clitics” in Korean linguistics (cf. Zingler(2022))
 - Against clitic approach (Chae (2020: 121f.): dependent nouns are obligatory leaners, a subtype of clitics)
 - **Clitic (접어)**: Prototypical manifestation, which combines a syntactic distribution with some degree of prosodic dependence on a host. (격조사, 복수표지, -스럽- 등)
 - **Anti-clitic (반접어)**: **Affixes** except that they share some behaviour with **phonological words** (것_{event-n-anti-clitic}, 것_{theme-n-anti-clitic}, 것_{theme-n-anti-clitic}, 놈_{argument-n-anti-clitic}, 놈_{event-n-anti-clitic}, 자_{agent-n-anti-clitic} 등)
- ② Three types of 것 from a morpho-syntactic perspective:
 - (의존)명사 = *noun-lexeme*: 것_{noun-lxm}
 - 보문소 = *comp-lexeme*: 것_{comp-lxm}
 - 반접어 = *anti-clitic*: 것_{event-n-anti-clitic}, 것_{theme-n-anti-clitic}, 것_{instrument-n-anti-clitic}

1. Objectives, Issues, and Basic Ideas [6/6]

- ③ Three types of 것 from a morpho-syntactic perspective:
 - (의존)명사 = *noun-lexeme*: 것_{noun-lxm}
 - 보문소 = *comp-lexeme*: 것_{comp-lxm}
 - 반접어 = *anti-clitic*: 것_{event-n-anti-clitic}, 것_{theme-n-anti-clitic}, 것_{instrument-n-anti-clitic}
- ④ Further classification of anti-clitic: 것_{event-n-anti-clitic}, 것_{theme-n-anti-clitic}, 것_{instrument-n-anti-clitic}
 - 것_{event-n-anti-clitic}: 사건 명사화 반접어(Event Nominalization Anti-Clitic, cf. *destruction*)
 - 것_{theme-n-anti-clitic}: 대상 명사화 반접어(Theme Nominalization Anti-Clitic, cf. *examinee*)
 - 것_{instrument-n-anti-clitic}: 도구 명사화 반접어(Instrument Nominalization Anti-Clitic, cf. *mixer*)
- ⑤ Differences and Similarities between 것, 놈, and 자
 - They are all anti-clitics.
 - 놈_{event-n-anti-clitic}: Event Nominalization Anti-Clitic
 - 놈_{argument-n-anti-clitic}: Argument(e.g. Agent/Theme/Instrument/...) Nominalization Anti-Clitic
 - 자_{human-arg-n-anti-clitic}: Argument[+HUMAN] Nominalization Anti-Clitic

2. Previous Analyses and Desiderata

2. Previous Analyses and Desiderata [1/3]

✓ Traditional/Reference Grammar

- Dependent or independence?
 - Previously: Incomplete noun, Defective noun (Sohn(1999))
 - Nam & Ko (1985: 74ff.): Dependent noun

🗨 Desiderata

- 전통적인 9 품사 중 명사로 보지만, 자립성에 방점을 두어 자립명사가 아니라는 의미로 의존명사라 칭함
- '것'의 형태-통사론적 분석이나 세분될 필요성에 대해서는 언급이 없음
- '것'이 단어(word)인지, 접사(affix)인지, 접어(clitic)인지, 그것도 아니면 반접어(anti-clitic)인지 형태-통사론적 논의 없음

3. Proposal: 것 as Anti-Clitic

3.1 Theoretical Preliminaries [1/2]

- ✓ Clitics vs. Affixes: Differences between word-clitic combinations and stem-affix combinations (cf. Stump 1998: 19–21; Spencer & Luis 2012: 108–111):
 - **Host selectivity**: Clitics can exhibit a low degree of selection with respect to their hosts, while affixes exhibit a high degree of selection with respect to their stems.
 - **Arbitrary gaps** in the set of combinations are more characteristic of affixed words than of clitic groups.
 - **Morphophonological idiosyncrasies** are more characteristic of affixed words than of clitic groups.
 - **Semantic idiosyncrasies** are more characteristic of affixed words than of clitic groups.
 - **Clitic-affix ordering**: Clitics can attach to material already containing clitics, but affixes cannot.

- (10) a. the boy^s play in the yard
 b. the boy^s [who live next door] play in the yard
 c. *the boy [who live next door]^s play in the yard
- (11) a. Simple clitics: I'd go, he'll leave, I can't bear, etc.
 b. Special clitics: [the man standing over there]'s hat

3.1 Theoretical Preliminaries [2/2]

✓ Clitics vs. Anti-clitics (Zingler (2022), cf. Chae(2020))

- **Clitic** (접어): Prototypical manifestation, which combines a syntactic distribution with some degree of prosodic dependence on a host. (격조사, 복수표지 등)
- **Anti-clitic** (반접어): **Affixes** except that they share some behaviour with **phonological words**. ('것', '놈', '자' 등)
- **Weak words** subsumes elements that are independent words in every regard except that they do not bear stress and/or tone.
- **"mobile" and "suspended" affixes**, which show types of syntagmatic freedom not found with prototypical affixes.
- 반접어(anti-clitics)란 접어(clitics)의 반대개념으로, 음운적으로는 독립성이 있지만 문법적으로는 독립성이 없는 표현
 - 음운적으로도 문법적으로도 독립성이 있는 표현은 우리가 일반적으로 알고 있는 단어(예: 대학교)
 - 음운적으로도 문법적으로도 독립성이 없는 표현은 접사(affix)(예: 명사화파생접미사 음/ㅁ/기).
 - 음운적으로는 독립성이 없지만 문법적으로는 독립성이 있는 표현은 접어(clitics)(예: 복수형태소 들, -스럽-, -답-, 접속형태소 다/라고).
 - 자립적으로 문장에 등장할 수 있기에 음운론적으로 단어이지만 문법적으로는 어휘소가 아닌 넓은 의미의 단어에 속한 표현들을 말한다. 즉, 독립성은 있지만 기능적으로 접사인 표현들은 반접어(anti-clitics) (Zúñiga (2014), Zingler (2022) 참고)

3.2 것_{noun-lxm}: Noun [1/1]

- ① 것_{noun-lxm} is a noun of type *spr-noun-lxm* that requires a determiner or NP[*gen*] as its specifier value.

- (12) a. [_{NP} *({이, 그, 저, 새, 헌, 옛, ...}) 것_{noun-lxm}]이 건강에 좋다.
b. 그 회사는 [_{NP} *(아버지의) 것_{noun-lxm}]입니까?

- ② A lexical entry for 것_{noun-lxm}

- (13)
- $$\left\langle \text{것}_{\text{noun-lxm}}, \left[\begin{array}{l} \text{ARG-ST} \left\langle \text{Det} \mid \text{NP}[\text{CASE } \textit{gen}] \right\rangle \\ \text{SEM} \left[\begin{array}{l} \text{INDEX} \left\langle \begin{array}{l} i \\ \text{RELN} \left\langle \begin{array}{l} \text{INST} \\ \text{thing} \end{array} \right\rangle \end{array} \right\rangle \end{array} \right] \end{array} \right] \right\rangle$$

3.3 것_{comp}: Complementizer [1/1]

- ① 것_{comp-lxm} is a lexeme of type *comp-lxm* that subcategorizes for a clause and marks clause boundaries.

- (14) a. 부모들은 [[아이들이 밥을 많이 먹는] (*새/*이/*어느/*몇/*한) 것_{comp-lxm}]]을 좋아하신다/몰랐다/알았다.
b. 철수가 [[적군이 도시를 파괴한] (*새/*이/*어느/*몇/*한) 것_{comp-lxm}]]을 보고하였다/몰랐다/알았다.

- ② A lexical entry for 것_{comp-lxm}

$$(15) \quad \left\langle \text{것}_{\text{comp-lxm}}, \begin{bmatrix} \text{comp-lxm} \\ \text{ARG-ST} \left\langle \left[\text{FORM adnom} \right] \right\rangle \\ \text{SEM} \left[\text{MODE prop} \right] \end{bmatrix} \right\rangle$$

3.4 것_{event-n-anti-clitic}: Event Nominalization Anti-Clitic [1/3]

- ① 것_{event-n-anti-clitic} is an anti-clitic of type *event-n-anti-clitic*,
- which subcategorizes for a clause and nominalizes the entire clause syntactically.
 - Its part of speech is a noun N.
 - The anti-clitic 것_{event-n-anti-clitic} is a phonological word that corresponds to the nominalization suffixes 'ㅁ/음/기', and
 - is similar to the gerundive derivational suffix *-ing* in English or *-ung* in German.
 - '것'은 명사화소이다. 명사화소로서 것_{event-n-anti-clitic}는 독자적인 지시체를 가지고 있지 않다.
 - 의미는 명제이고, 명제의 의미 내용은 것_{event-n-anti-clitic}이 하위범주화하는 절의 내용과 동일하다.

- ② A lexical entry for 것_{event-n-anti-clitic}

$$(16) \quad \left\langle \text{것}_{\text{event-n-anti-clitic}}, \begin{bmatrix} \text{event-n-anti-clitic} \\ \text{ARG-ST} \left\langle \left[\begin{array}{c} \text{FORM adnom} \\ \text{INDEX } s \end{array} \right] \right\rangle \\ \text{SEM} \left[\begin{array}{c} \text{MODE prop} \\ \text{INDEX } s \end{array} \right] \end{bmatrix} \right\rangle$$

3.4 것_{event-n-anti-clitic}: Event Nominalization Anti-Clitic [2/3]

- ③ 품사가 명사인 증거: '때문'의 선행 명사로 가능 & NP 논항을 하위범주화하는 술어의 논항 자리에 등장

- (17) a. 사람들은 [너무 많이 먹기] 때문에 살이 쪼든다.
b. 사람들은 [너무 많이 먹는 것_{event-n-anti-clitic}] 때문에 살이 쪼든다.
- (18) a. $\left\{ \begin{array}{l} \text{[NP 아이]가} \\ \text{[NP 아이가 웃는 것}_{\text{event-n-anti-clitic}}\text{]}이 \end{array} \right\}$ 귀엽다.
b. 철수는 $\left\{ \begin{array}{l} \text{[NP 컴퓨터]를} \\ \text{[NP 컴퓨터가 고장난 것}_{\text{event-n-anti-clitic}}\text{]}을 \end{array} \right\}$ 고쳤다.

3.4 것_{event-n-anti-clitic}: Event Nominalization Anti-Clitic [3/3]

- ④ 것_{event-n-anti-clitic}은 용언(predicate)을 명사로 만드는 파생 반접어로서 논항의 실현 여부에 대한 제약을 두지 않는다.

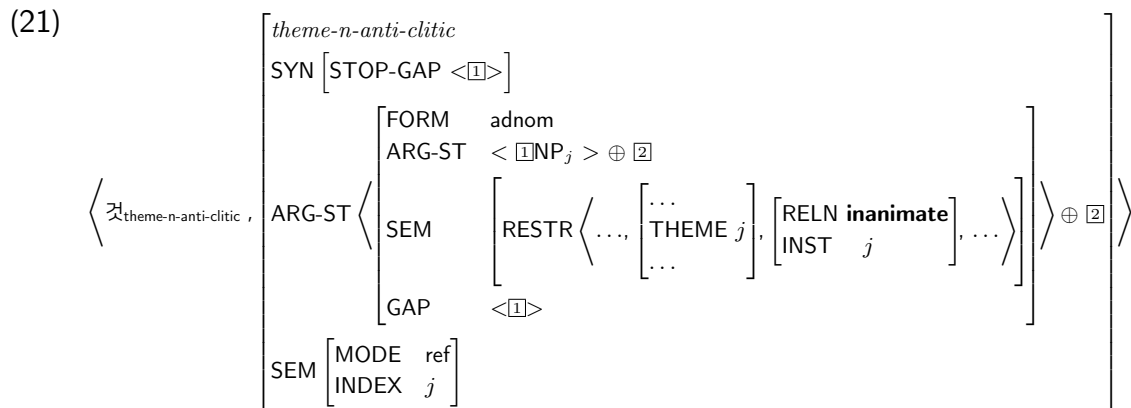
- (19) a. 나는 [CP 먹는 것_{event-n-anti-clitic}]이 좋아.
b. 철수가 [CP 고양이가 쥐를 쫓는 것_{event-n-anti-clitic}]을 둘 다 잡아버렸다.

- ⑤ 것_{event-n-anti-clitic} allows multiple antecedents in the constructional constraint of type *event-for-participant-cx* (Ryu(2023) 참고).

- (20) a. 철수는 [[고양이가 쥐를 쫓고 있는] 것_{event-n-anti-clitic}]을 둘 다 잡았다. ('고양이와 쥐')
b. 철수는 [[고양이가 쥐를 쫓고 있는] 것_{event-n-anti-clitic}]을 둘 다 보았다. ('고양이와 쥐')

3.5 것_{theme-n-anti-clitic}: Theme Nominalization Anti-Clitic [1/3]

- ① An anti-clitic of type *theme-n-anti-clitic* that
 - Subcategorizes for a verbal lexeme and syntactically nominalizes a non-agent argument of the clause, and
 - Its part of speech is a noun N.
- ② A lexical entry for 것_{theme-n-anti-clitic}



3.5 것_{theme-n-anti-clitic}: Theme Nominalization Anti-Clitic [2/3]

- ③ 독일어의 -er, -or, 영어의 -ee, -ling 명사화에서 보이는 경우와 동일 (Spencer (2005: 87) 더 자세한 논의는 Baker (1988) 참고)

- (22)
- a. -ee in English (e.g., *employee*)
 - b. -ling in German (e.g., *Lehrling*)
 - c. employee, refugee, trainee, trustee, devotee, interviewee, lessee, licensee, nominee, referee, etc.
 - d. amputee, appointee, auditee, assignee, devotee, examinee, patentee, transferee, abductee, internee, etc.

- (23)
- a. [[____ 한국적인] 것_{theme-n-anti-clitic}]의 특징/설명/홍보 (예: K-Pop)
 - b. [[____ 어제 먹은] 것_{theme-n-anti-clitic}]의 레시피/찬평/지불 (예: 음식)

3.5 것_{theme-n-anti-clitic}: Theme Nominalization Anti-Clitic [3/3]

- ④ THEME을 만족시키는 반접어(anti-clitic)라는 증거: '하다'의 목적어에 할당되는 의미역, '되다'의 주어에 할당되는 의미역이 동일한 경우 '것_{theme-n-anti-clitic}'과 결합하여 동일한 의미를 가지게 됨

- (24) a. [[유명한] 것] = [[유행된] 것_{theme-n-anti-clitic}]
 b. [[도착한] 것] = [[도착된] 것_{theme-n-anti-clitic}]
 c. [[연구한] 것] = [[연구된] 것_{theme-n-anti-clitic}]
 d. [[쫓는] 것] = [[쫓긴] 것_{theme-n-anti-clitic}]

- ⑤ 수식절 안에 대상 명사가 실현되지 않아야 문법적이다 (cf. [STOP-GAP <1>]).

- (4) a. 철수는 [[영희가 무쇠절구통으로 다진] 것_{theme-n-anti-clitic}]을 넣고 김치찌개를 끓였다.
 b. *철수는 [[영희가 무쇠절구통으로 마늘을 다진] 것_{theme-n-anti-clitic}]을 넣고 김치찌개를 끓였다.

- ⑥ 내포절 술어와 관련된 어떠한 논항도 실현되지 않은 경우에 대상 해석이 가능하다.

- (5) 철수는 [[다진] 것_{theme-n-anti-clitic}]을 보관하였다. (대상: 마늘, 쇠고기, 돼지고기 등)

3.6 것_{instrument-n-anti-clitic}: Instrument Nominalization Anti-Clitic [1/4]

- ① 내핵관계절 가설에 따르면 도구 명사구는 명시적으로 실현되지 않았기에 (1)이 (2c)로 해석되는 현상은 분석할 수 없었으나 반접어로 분석하면 분석 가능

- (1) 철수는 [[영희가 마늘을 다진] 것]을 보관하였다.
 (2) c. 철수는 [영희가 마늘을 다진] 도구를 보관하였다. (도구)

- ② A lexical entry for 것_{instrument-n-anti-clitic}

- (25)
- $$\left\langle \text{것}_{\text{instrument-n-anti-clitic}}, \begin{array}{l} \text{SYN} \left[\text{STOP-GAP} \langle 1 \rangle \right] \\ \text{ARG-ST} \left\langle \begin{array}{l} \text{FORM} \quad \text{adnom} \\ \text{ARG-ST} \quad \langle 1 \text{NP}_k \rangle \oplus 2 \\ \text{SEM} \quad \left[\text{RESTR} \left\langle \dots, \left[\dots \text{INSTRUMENT } k \dots \right] \right\rangle \right] \oplus 2 \\ \text{GAP} \quad \langle 1 \rangle \end{array} \right\rangle \\ \text{SEM} \left[\begin{array}{ll} \text{MODE} & \text{ref} \\ \text{INDEX} & k \end{array} \right] \end{array} \right\rangle$$

3.6 것_{instrument-n-anti-clitic}: Instrument Nominalization Anti-Clitic [2/4]

- ③ 것_{instrument-n-anti-clitic} 이 도구로 해석되는 경우 수식절 안에 도구 명사가 실현되지 않으면 문법적이나 오히려 모두 실현되면 비문법적임을 잘 설명 (cf. [STOP-GAP <①>]).

- (3) a. 철수는 [[영희가 마늘을 다진] 것]으로 돼지고기도 다졌다. (도구)
b. *철수는 [[영희가 무쇠팔구통으로 마늘을 다진] 것]으로 돼지고기도 다졌다. (도구)

- ④ 내포절 술어와 관련된 어떠한 논항도 실현되지 않은 경우에 도구 해석과 대상 해석은 가능해서 (5)는 중의적이다.

- (5) 철수는 [[다진] 것]을 보관하였다.
a. 대상: 마늘, 쇠고기, 돼지고기 등
b. 도구: 마늘다지기, 무쇠팔구, 핸드블랜더, 갈릭프레스, 믹서, 식칼 등

3.6 것_{instrument-n-anti-clitic}: Instrument Nominalization Anti-Clitic [3/4]

- ⑤ 것_{theme-n-anti-clitic} 과 것_{instrument-n-anti-clitic} 의 경우 실현되지 않은 넓은 의미의 확대 논항 구조를 가정하고 그 논항구조 내부의 한 논항을 실현시키는 것으로 이해한다. 이때 중의성이 존재하나 해석의 선호도는 있는 것으로 보인다. 일반적으로 대상 → 도구의 순이나, 반복성 여부에 따라 도구 → 대상 해석이 선호된다..

- (26) a. 나는 [_{NP} [_S ____ ____ 다진] 것]을 사왔다.
해석 선호도: 것_{theme-n-anti-clitic} → 것_{instrument-n-anti-clitic}
b. 나는 [_{NP} [_S ____ ____ 다지는] 것]을 사왔다.
해석 선호도: 것_{instrument-n-anti-clitic} → 것_{theme-n-anti-clitic}

3.6 것_{instrument-n-anti-clitic}: Instrument Nominalization Anti-Clitic [4/4]

- ⑥ 대상(THEME) 명사가 선행절 내부에 이미 실현되어 있으면 일반적으로 것_{theme-n-anti-clitic} 이 아니지만, 문장은 중의적일 수 있다. 대상 의미역이 실현되어 있으면 한편으로는 다중격표지구문이 되거나 아니면 도구(INSTRUMENT) 의미역을 지칭하는 것으로 이해가 된다. 전자는 것_{theme-n-anti-clitic} 이고 후자는 것_{instrument-n-anti-clitic} 이다.

(27) a. 나는 [_{NP} [_S ____ 마늘을 다진] 것]을 사왔어.

중의적: (1) 다진 마늘 = 다중격표지 = 것_{theme-n-anti-clitic}

(2) 도구/물품 = 것_{instrument-n-anti-clitic}

b. 나는 [_{NP} [_S 철수가 마늘을 다진] 것]을 사왔어.

중의적: (1) 철수가 다진 마늘 = 다중격표지 = 것_{theme-n-anti-clitic}

(2) 도구/물품 = 것_{instrument-n-anti-clitic}

4. Similarities and Differences Between 것, 놈, and 자

4. Similarities and Differences between 것, 놈, and 자 [1/4]

① Similarities and Differences between 것, 놈, and 자

- (6) a. 철수는 [[영희가 마늘을 다진] { 것을/놈을/*자를}] 찾았다.
 b. 철수는 [[마늘을 다진] { 것을/놈을/자를}] 찾았다.
 c. 철수는 [[다진] { 것을/놈을/자를}] 찾았다.
- (7) a. 다진 $\left\{ \begin{array}{l} \text{것} (= \text{행위}, = \text{대상}, = \text{도구}, \neq \text{행위자}) \\ \text{놈} (\neq \text{행위}, = \text{대상}, = \text{도구}, = \text{행위자}) \\ \text{자} (\neq \text{행위}, \neq \text{대상}, \neq \text{도구}, = \text{행위자}) \end{array} \right\}$
 b. 마늘을 다진 $\left\{ \begin{array}{l} \text{것} (= \text{행위}, = \text{대상}, = \text{도구}, \neq \text{행위자}) \\ \text{놈} (\neq \text{행위}, \neq \text{대상}, \neq \text{도구}, = \text{행위자}) \\ \text{자} (\neq \text{행위}, \neq \text{대상}, \neq \text{도구}, = \text{행위자}) \end{array} \right\}$
 c. 영희가 마늘을 다진 $\left\{ \begin{array}{l} \text{것} (= \text{행위}, \neq \text{대상}, = \text{도구}, \neq \text{행위자}) \\ \text{놈} (\neq \text{행위}, \neq \text{대상}, = \text{도구}, \neq \text{행위자}) \\ \text{*자} (\neq \text{행위}, \neq \text{대상}, \neq \text{도구}, \neq \text{행위자}) \end{array} \right\}$

4. Similarities and Differences between 것, 놈, and 자 [2/4]

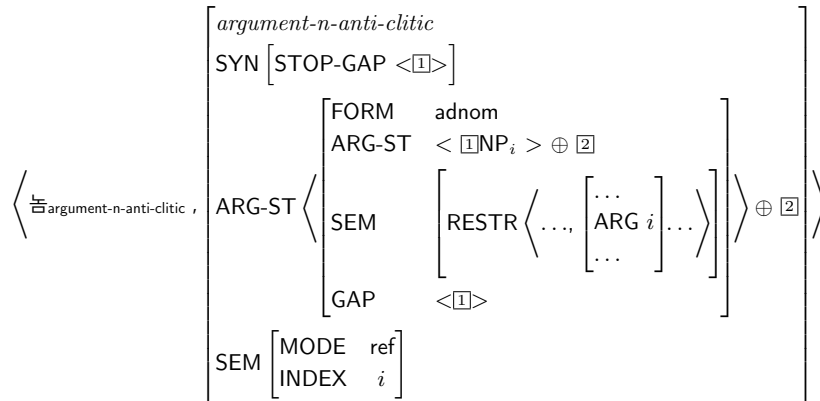
② A lexical entry for 자_{human-arg-n-anti-clitic}

- (28)
- $$\left\langle \text{자}_{\text{human-arg-n-anti-clitic}}, \left[\begin{array}{l} \text{argument-n-anti-clitic} \\ \text{SYN} [\text{STOP-GAP} \langle \text{1} \rangle] \\ \text{ARG-ST} \left\langle \begin{array}{l} \text{FORM} \quad \text{adnom} \\ \text{ARG-ST} \quad \langle \text{1} \text{NP}_i \rangle \oplus \text{2} \\ \text{SEM} \quad \left[\text{RESTR} \left\langle \dots, \left[\begin{array}{l} \dots \\ \text{ARG } i \end{array} \right], \left[\begin{array}{l} \text{RELN human} \\ \text{INST } j \end{array} \right], \dots \right\rangle \right] \oplus \text{2} \\ \text{GAP} \quad \langle \text{1} \rangle \end{array} \right\rangle \\ \text{SEM} \left[\begin{array}{l} \text{MODE ref} \\ \text{INDEX } i \end{array} \right] \end{array} \right] \right\rangle$$

4. Similarities and Differences between 것, 놈, and 자 [3/4]

3 A lexical entry for $\text{놈}_{\text{argument-n-anti-clitic}}$

(29)

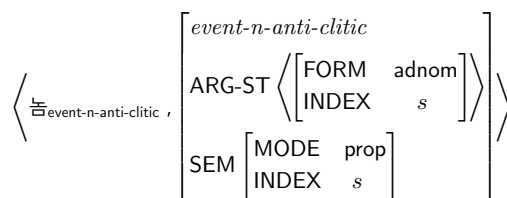


4. Similarities and Differences between 것, 놈, and 자 [4/4]

4 A lexical entry for $\text{놈}_{\text{event-n-anti-clitic}}$

- (30) a. 경찰이 [[도둑이 은행에서 나오는] 놈]을 현장에서 체포하였다.
 b. 철수가 [[컴퓨터가 고장난] 놈]을 고쳤다.
 c. 할머니께서 [[아이가 우는] 놈]을 간신히 달랬다.

(31)



5. Conclusion

5. Conclusion [1/2]

i 기본 아이디어

- '것'과 선행절 결합체의 해석에 논항의 실현 여부는 상관 없이, 기본적으로 선행절 술어의 논항구조와 관련되어 있다.
- '것'은 이들 중 특정 논항을 실현시켜 명사로 파생시키는 기능을 담당한다 (cf. *destruction*, *examinee*, *mixer*).
- 이 명사 파생 기능은 파생 접사(affix)로서도 아니고 음운론적 자립 단어(word)로서도 아니고 반접어(anti-clitic)로서 수행한다.
- 것, 놈, 자는 반접어(anti-clitic)로서 음운론적으로 독립적인 단어이나 역할은 접사와 같이 명사 파생 역할을 한다는 아이디어를 증거와 함께 장점을 제시

5. Conclusion [2/2]

✓ 결론 및 의의

- ① Introduction of the notion of “anti-clitics” in Korean linguistics
- ② Against clitic approach (Chae (2020: 121f.): dependent nouns are obligatory leaners, a subtype of clitics)
- ③ Three types of 것 from a morpho-syntactic perspective:
 - (의존)명사 = *noun-lexeme*: 것_{noun-lxm}
 - 보문소 = *comp-lexeme*: 것_{comp-lxm}
 - 반접어 = *anti-clitic*: 것_{event-n-anti-clitic}, 것_{theme-n-anti-clitic}, 것_{instrument-n-anti-clitic}
- ④ Further classification of anti-clitic: 것_{event-n-anti-clitic}, 것_{theme-n-anti-clitic}, 것_{instrument-n-anti-clitic}
 - 것_{event-n-anti-clitic}: 사건 명사화 반접어(Event Nominalization Anti-Clitic, cf. *destruction*)
 - 것_{theme-n-anti-clitic}: 대상 명사화 반접어(Theme Nominalization Anti-Clitic, cf. *examinee*)
 - 것_{instrument-n-anti-clitic}: 도구 명사화 반접어(Instrument Nominalization Anti-Clitic, cf. *mixer*)
- ⑤ Differences and Similarities between 것, 놈, and 자
 - They are all anti-clitics.
 - 놈_{event-n-anti-clitic}: Event Nominalization Anti-Clitic
 - 놈_{argument-n-anti-clitic}: Argument(e.g. Agent/Theme/Instrument/...) Nominalization Anti-Clitic
 - 자_{human-arg-n-anti-clitic}: Argument[+HUMAN] Nominalization Anti-Clitic

Selected References [1/2]

- Chae, Hee-Rahk (2020), *Korean Morphosyntax: Focusing on Clitics and Their Roles in Syntax*, London and New York: Routledge.
- Chung, Chan (2000), Complex Noun, Multiple Inheritance, and Internally Headed Relativization in Korean, *Proceedings of the 25th Annual Meeting of the Berkeley Linguistics Society*, Vol.25, 50–61.
- Halpern, Aaron (1998), Clitics. In Andrew Spencer & Arnold M. Zwicky (eds.) (1998), *The Handbook of Morphology*, 101-122, Oxford: Blackwell.
- Jhang, Sea-Eun (1994), *Headed Nominalizations in Korean: Relative Clauses, Clefts, and Comparatives*, Ph.D. thesis, Simon Fraser University.
- Kim, Jong-Bok (2016), *The Syntactic Structures of Korean: A Construction Grammar Perspective*, Cambridge: Cambridge University Press.
- Kim, Min-Joo (2004), *Event-Structure and the Internally-Headed Relative Clause Construction in Korean and Japanese*, Ph.D. thesis, University of Massachusetts at Amherst.

Selected References [2/2]

- Kim, Yong-Beom (2002), Relevancy in Internally Headed Relative Clauses in Korean, *Lingua* 112, 541–559.
- Song, Sanghoun (2021), Different Types of Internally Headed Relative Clauses in Korean: A Corpus-based Analysis, *Studies in Linguistics* 60, 89–114.
- Spencer, Andrew & Arnold M. Zwicky (eds.) (1998), *The Handbook of Morphology*, Oxford: Blackwell.
- Yoon, Hang-Jin (2012), Different Types of ‘것’ Constructions in Korean, *Studies in Generative Grammar*, Volume 22, Number 3, 557–577.
- Zúñiga, Fernando (2014), (Anti-)cliticization in Mapudungun, *Morphology* 24, 161–175.
- 류병래 (2022), 한국어 내핵 관계절의 존재에 대한 증거와 비증거, 『언어와 정보』 26(2), 1–25.
- 류병래 (2023), Korean Internally-Headed Relative Clauses as “EVENT FOR PARTICIPANT” Metonymy, 한국언어정보학회·대한언어학회·한국현대언어학회·언어과학회·현대문법학회 2023 공동학술대회 발표논문집 (Sat., May 20, 2023, 계명대학교).

CLAUSAL ELLIPSIS AND NON-SIMULTANEOUS TRANSFER

Haewon Jeon (KMOU)

I . What's Clausal Ellipsis and Why Should We Care?

- ★ a **mismatch** between sound and meaning.
- ★ Clausal Ellipsis

II. The Internal Structure of the Ellipsis site

- ★Fragments
- ★ Remnants
- ★Head as **ellipsis licensor**
- ★ Structure of the CP: Two Head Positions
- ★ Duality of Semantics (Chomsky, 2020)

III. How and When to Transfer

- ★ Non-Simultaneous Transfer
- ★Variability of Transfer system
- ★ Mismatches between PHON and SEM

IV. Conclusion

- ★ How do we model these?

ROAD MAP

I . What's Clausal Ellipsis and Why Should We Care?

★ Ellipsis is a mismatch between sound and meaning. In other words, **certain syntactic structures(elliptical sites) are not met in the phonetic realization of the sentence.**

(1) John can play something, but I don't know what ~~John can play~~. **Clausal-Ellipsis** (Merchant 2013:3)

Chomsky and Lasnik (1993: 564) assume that ellipsis is an extreme form of **phonological reduction** in a typical manner. (see also Rooth, 1992; Tancredi, 1992; Vanden Wyngaerd and Zwart, 1999; Hartmann, 2000). Merchant (2001, 2004) argues that a fragment answer and sluicing are derived by **syntactic A'-movement of the remnant and deletion of TP**. As is well known, this movement and deletion approach sheds light on ellipsis is a genuinely syntactic phenomena, not deaccented material.

I . What's Clausal Ellipsis and Why Should We Care?

★ Clausal Ellipsis : Fragments have **a hidden structure** based on *Movement+Deletion analysis* (Merchant, 2004). Ellipsis is a type of discourse anaphor. Just like other discourse anaphora (such as pronouns), it takes its referent from an antecedent.

(2) a. A: What did she eat? B: She ate **beans**. **Fragment Answer**

b. A: What did she eat?

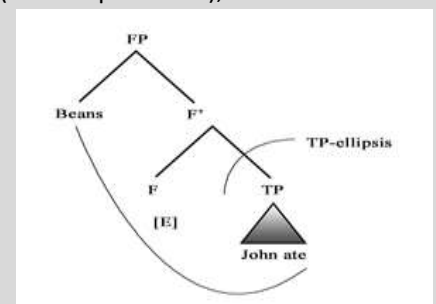
B: **Someone ate beans, but I don't know who** ate beans. **sluicing**

c. A: 철수는 무엇을 먹었니? B: 빵

A': 혜원은? [**Fragmentary Question in Korean** 혜원은 무엇을 먹었니]

d. Which tree did Juan plant, **the oak** Juan planted?

Split Question



★ **Goals: The quest for a better understanding of the nature of ellipsis + in current Minimalist theorizing.**

II. The Internal Structure of the Ellipsis site

The prevailing idea about clausal ellipsis is that it involves the deletion of the **periphery of the elliptical clause**.

Wei (2013) assumes that **Fragmentation** is a property of ellipsis that remnant properties that remnant move up to the ellipsis site.

Maeda (2019) and Maeda and Jeon (2019) postulate that the internal structure of remnant movement and clausal deletion.

Sluicing in English also applies to Korean, involving focus movement that of FAs in English, involving focus movement.

Maeda and Jeon (2019) and Jeon (2022ab) postulate that **topic that can survive from ellipsis** is a property of ellipsis that remnant properties that remnant move up to the ellipsis site.

★ Remnants: Ellipsis of a clause to the exception of a single constituent

(4) A: Chelswu-nun Yenghuy-eykey mwues-ul cwess ni?

Chelswu-top Yenghuy-dat what-acc gave q

‘What did Chelswu give to Yenghuy?’

B: Wine ipnida.

wine cop

‘Wine.’

A: So-Yun-eykey-nun? (FQ)

So-Yun-dat-top

To So-Yun? (Intended: What did Chelswu give to So-Yun?)’

(3) [**Remnant**]_{FocP} (wh) Foc [wh [TP]] (Maeda and Jeon 2019:174)

II. The Internal Structure of the Ellipsis site

Maeda and Jeon(2019) postulate that **a contrastive topic remnant undergoes overt movement to Top-FocP**, and a wh-element undergoes overt/covert movement to FocP, **in a cartographic left periphery**

(5) [**ForceP** Force [**TopP** Top [**FocP** Foc [**TopP** Top [**FinP** Fin [TP...]]]] (Rizzi, 1997: 297)

Given the syntactic properties of FQs in Korean (Maeda and Jeon 2019, Jeon 2022ab), ellipsis constructions if it is elided, need all information of syntactic objects **including A-bar position as focus or topicalization**. Especially, in FQs in Korean, remnants in the left periphery of the clause are required to be contrastively topicalized in A-bar position as schematized above.

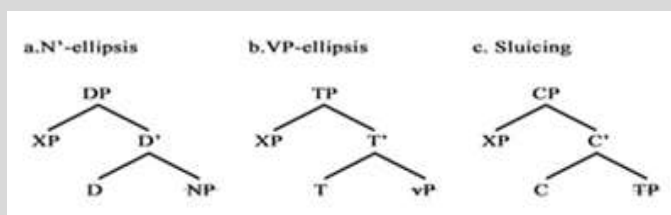
II. The Internal Structure of the Ellipsis site

★ Head as ellipsis licenser

(6) a. [vP **V°**[E] [VP ...]] (VP-deletion)

b. [TP **T°**[E] [vP [VP ...]]] (vP-deletion)

c. [CP **C°**[E] [TP [vP [VP ...]]]] (TP-deletion) (Lobeck 1991, 1995)



Saito et al (2009)

A functional head H mediating the remnant (Contrastive Topic) and elided material in FQs in Korean armed with den Dikken (2008), and Ko (2011). See also Park(2012a, 2012b, 2013a). She assumed functional heads VCAUS, VBE, VPASS as a phase head. More on a functional head, refer to AspP in Dutch (Aelbrecht 2010), the CP in English and German (Abels 2018b), and Pre P of Miyagawa (2001).

II. The Internal Structure of the Ellipsis site

★ Structure of the CP: Two Head Positions

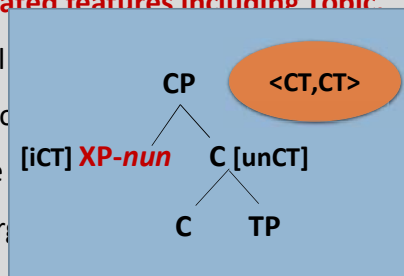
(7) [CP1 **XP-nun** [CP2 C [**TP** ...XP]]

Based on these generalizations, Jeon(2019,2022b) assumed that there exists a functional head H mediating the remnant(Contrastive Topic) and elided material in FQs in Korean.

We pointed out that FQs are derived by fronting the target to the specifier position of CP2, a functional head in the left periphery higher than TP projection. The target of fragments is new information, in contrast to the old information in the antecedent.

II. The Internal Structure of the Ellipsis site

Jeon(2022b) assumed that **the higher CP hosts the discourse related features including Tonic** while **the lower CP concerns the clausal typing features**. We will **data on Fragments**. In the spirit of movement and deletion approach, Korean are generated from the alternative checking of discourse related features. Jeon postulates that contrastively topic-marker remnants undergo target deletion. The higher CP encodes discourse related interpretable features, and CP has its uninterpretable counterpart. what is coded by <CT, CT> can be a remnant, which can survive from the clausal deletion.



II. The Internal Structure of the Ellipsis site

★ Clausal Ellipsis

(8) a. Sluicing: [CP₁ **XP** [CP₂ C [TP ... XP ...]]

b. FA(fragment Answer): [CP₁ **XP** [CP₂ C [TP ... XP ...]]

c. FQ(fragmentary Question) : [CP₁ **XP** [CP₂ C [TP ... XP ...]]

d. SQ(split question) : [CP₁ **XP** [CP₂ C [TP ... XP ...]]

(9) [**CP**₁ **XP** [**CP**₂ C [TP ... **XP**]

Phonetic Realization (o) Phonetic Realization (x)

Semantic Interpretation (o) Semantic Interpretation (o)

★ Duality of Semantics (Chomsky, 2020)

The notion, Duality of Semantics (Chomsky, 2020), that CI incorporates a dual semantics, with generalized argument structure as one component, the other being discourse-related and scopal properties originates from Chomsky (2007: 5, 2008: 7) that **External Merge(EM)** yields argument structure(agent, patient, goal, etc.), and **Internal Merge(IM)** yields discourse-related properties such as topic, old information and specificity, along with scopal effects.

III. How and When to Spell-out/Transfer

(10) [CP₁ XP [CP₂ C [TP ...XP]] *(a unified version of Clausal Ellipsis, Jeon(2022b))*

As briefly discussed above, we provided a novel way to shed light on the licensing head of the ellipses in sentence fragments such as Sluicing, FA(fragment answer), FQ(fragmentary Question), SQ(split question) in a unified fashion.

Then, a crucial question is the timing of Transfer. Is structural information sent to the two interfaces at the same time or do the two interfaces receive different information at different times? Given this situation, the question is why there exist different loci of transferring resultant labels to both interfaces. The question remains, however, how parallel or independent are the transferences to CI and SM?

vP and CP are the only loci to transfer?

The literature has put forth a number of analyses for the problem of how a syntactic derivation proceeds. When words are combined to form a sentence at the **D-Structure**, they are mapped to the **S-Structure** through various operations, such as movement, and then they are mapped to **the two interfaces, namely Logical Form and Phonetic Form (Chomsky,1981)**. The way a derivation proceeds has altered, though, according to Chomsky (1995), who abandoned both the D-Structure and the S-Structure. A derivation is now separated into several chunks. **Chomsky (2000, 2001) call this chunk “phase”**. Chomsky assumes that **vP** and **CP** are phases, and that a derivation proceeds phase by phase. →**“Transfer.”**

According to Chomsky, the labels or info of syntactic objects are transferred synchronously to the two interfaces. However, we should speculate this system can be generalized for human languages.

If PF and LF are timed differently → **Felser 2004, Matushansky 2005, Citko (2014) and Marušič (2005, 2009)**

- (13) **Transfer** applies to the **complement of a phase head** and maps it to **interfaces**.
(Chomsky, 2000)

As pointed out by Obata(2010) and Goto(2011), there is a theoretical problem in the definition in (12).

- (14) a. What did you eat?
b. [CP what did [TP you T [vP v [VP eat ___]]]

★ Mismatch between PHON and SEM

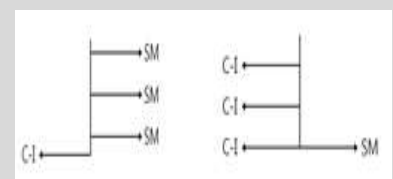
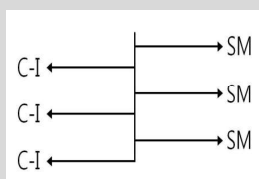
The two most obvious phenomena where the locus of interpretation differs from the locus of pronunciation: total reconstruction and quantifier raising. Namely, if at some point in the derivation, the structure built thus far gets spelled-out to one but not to the other interface, whatever was meant for the other interface but was not spelled-out to this interface remains in the derivation. As a result, the position of an element would differ from one interface to the other.

III. How and When to Spell-out/Transfer

★ Non-Simultaneous Spell-out

Citko (2014) claims that Non-Simultaneous Multiple Transfer is well designed to handle a situation in which a given element is phonologically visible in one position but semantically visible in another position.

Citko (2014) and Marušič (2005, 2009) argue that in the case of reconstruction, Spell-Out to the LF interface occurs before the Spell-Out to the PF interface. However, QR is the opposite. Namely, in the case of QR, Spell-Out to the PF interface occurs before the Spell-Out to the LF interface. Under Marušič's Transfer system, Transfer happens at different loci in the derivation. Marušič (2005, 2009) also argues that covert movement such as QR involves Transfer to PF before Transfer to LF.

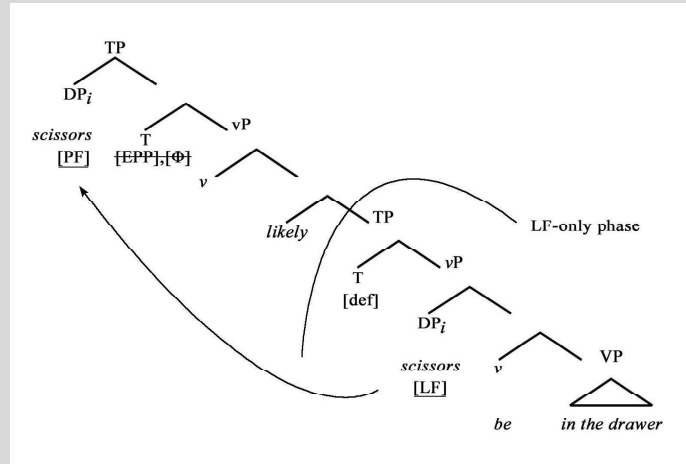


If the same syntactic objects are transferred synchronously to the two interfaces,
How did they communicate between SM and CI?

Non-Simultaneous Transfer

Marušič (2005, 2009)

(11) Scissors are likely to be in the drawer.



★ How do we model these?

(12) a. Sluicing: [cp1 **XP** [cp2 C [TP...XP...]]

b. FA(fragment Answer): [cp1 **XP** [cp2 C [TP...XP...]]

c. FQ(fragmentary Question) : [cp1 **XP** [cp2 C [TP...XP...]]

d. SQ(split question) : [cp1 **XP** [cp2 C [TP...XP...]]

Jeon(2022b)

e. Syntax: [cp1 **XP** [cp2 C [TP ...XP] (a unified version of Clausal Ellipsis)

SM : [cp1 **XP** _{Transfer2} [cp2 C [TP ...XP] _{Transfer 1}

CI : [cp1 **XP** [cp2 C [TP ...XP] Jeon(2023, to appear)

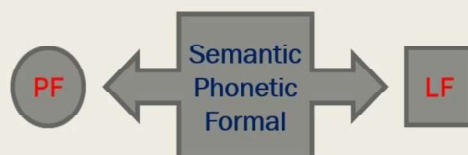
Viable Options:

The transferences to CI and SM would be parallel or independent ?

1) [CP₁ XP] [CP₂ C [TP ...XP]

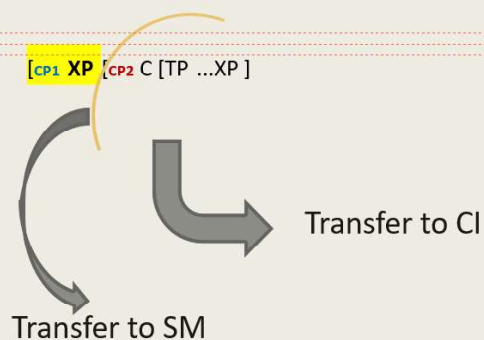


2) [CP₁ XP] [CP₂ C [TP ...XP]



Or 3) Every feature bundle goes to both interfaces?

How to encode 'Clausal Ellipsis' from Narrow Syntax to Interfaces



Works Cited (Selected)

- Aelbrecht, Lobke. 2010. *The syntactic licensing of ellipsis*. Amsterdam: John Benjamins Publishing.
- Citko, Barbara. 2014. *Phase Theory*. Oxford: Oxford University Press.
- Chomsky, Noam. 2008. On phases. *Foundational Issues in Linguistic Theory: Essays in Honor of Jean-Roger Vergnaud*. Robert Friedin, Carlos P. Otero, and Maria Luisa Zubizarreta, eds. Cambridge, MA: MIT Press.
- Chomsky, Noam. 2015. "Problems of Projection: Extensions." In E. Di Domenico, C. Hamann, and S. Matteini (eds.), *Structures, Strategies and Beyond: Studies in honour of Adriana Belletti*. John Benjamins. P.1-16.
- Chomsky, Noam. 2020. "Minimalism: where we are now, and where we are going." Talk at Linguistic Society of Japan. Sept. 22nd. Available at <https://www.youtube.com/watch?v=X4F9NSVVUw>.
- Lobeck, Anne. 1995. *Ellipsis: Functional Heads, Licensing, and Identification*, Oxford: OUP
- Marušič, Franc Lanko. 2005. "On Non Simultaneous Phases." Doctoral dissertation. StonyBrook University.
- Marušič, Franc Lanko. 2009. "Non-Simultaneous Spell-Out in the clausal and Nominal Domain." In *Interphases: Phase-Theoretic Investigations of Linguistic Interfaces*, ed. K.K. Grohmann, Oxford University Press.
- Maeda, Masako. 2019. "Fragmentary Questions and the Left Periphery in Japanese." *proceedings of 2nd FLC & NGC Joint Conference Pukyong National University*.

Works Cited (Selected)

- Maeda, Masako and Haewon, Jeon. 2019. "Fragmentary Questions in Japanese and Korean.", *proceedings of GLOW-in-Asia XII & SICOGG 21*, Dongguk University.
- Merchant, Jason. 2001. *The Syntax of Silence: Sluicing, Islands, and the Theory of Ellipsis*. Oxford: Oxford University Press.
- Merchant, Jason. 2005. "Fragments and Ellipsis." *Linguistics and Philosophy* 27, p.661-738.
- Merchant, Jason. 2013. "Diagnosing Ellipsis." In Cheng, L. and N. Cover, (eds.), *Diagnosing syntax*. Oxford: Oxford University Press. P.537-542.
- Jeon, Haewon. 2019. "Fragmentary Questions in Korean and Non simultaneous Transfer", *proceedings of NGC&FLC 3rd Joint Conference, Fukuoka National University*.
- Jeon, Haewon. 2022a. "Notes on Fragmentary Questions in Korean." *The journal of Humanities and Social Sciences* 23.2, p.117-138.
- Jeon, Haewon. 2022b "The Syntactic Nature of Remnants in Fragment Questions in Korean." *Journal of Language Sciences*, 29.3 p 111-132.

A Study on Sentence Stress and Isochrony Produced by Korean EFL Learners



연세대학교 석사과정 최아녕

목차

- I. 서론
- II. 배경
- III. 연구 질문
- IV. 데이터 및 연구 방법
- V. 결과
- VI. 결론 및 한계

서론

→ 제 2 언어 학습

- ◆ 제 1언어의 간섭 및 영향을 받음
- ◆ 음성 체계가 제 1 언어와 다를 경우 제 2 언어 학습자들이 어려움을 겪음
 - ◆ e.g., 한국인 영어 학습자

→ 모든 언어는 특정한 리듬을 가짐

- ◆ e.g., stress-timed, syllable-timed, and mora-timed

영어 vs. 한국어



stress-timed:

강세가 리듬적 시간
간격으로 나타남



syllable-timed:

음절이 일정한 시간
간격으로 나타남

배경

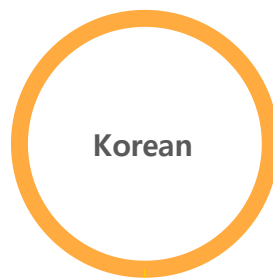
- 소리의 특징(sound feature)과 한국 영어 학습자의 능숙도 등급(proficiency level)
 - 소리의 음향적 특징들이 학습자의 숙련도 등급 구분에 어떤 역할을 하는 가?
 - ◆ 폐쇄음(stop), 마찰음(fricative) - 등급 별 차이가 있음
 - ◆ 기식(aspiration) 같은 경우에는 등급 구별에 차이를 크게 주지 않음
- 영어의 등시성(isochrony)
 - ◆ 영어 문장에서 강세(stress)는 일정한 간격을 두고 규칙적으로 반복되는 경향이 있음
 - ◆ e.g., My brother is coming on Friday

영어 vs. 한국어



stress-timed:

영어 문장에서 강세가
규칙적으로 반복



syllable-timed:

음절마다 비교적 일정
하게 강세를 받음

등시성(isochrony)이라는 영어의 특징이 한국인 영어 발화에서 어떻게 구현되며,
이는 학습자들의 영어 능숙도 등급(proficiency levels) 구별에 어느정도 역할을 할 수 있는 가?

데이터

→ 한국인의 영어 발음 음성 코퍼스 (Korean-Spoken English Corpus, K-SEC)

set	내용	단어, 문장 개수(개)		
		일반	중등	초등
set1	한국어 기본단어	98	98	98
set2	바람과 햇님	1	1	1
set3	영어자모음	64	64	64
set4	영어어휘	1125	956	435
★ set5	영어문장	★ 36	36	36
set6	바람과 햇님(영문)	1	1	1
합계		1325	1156	635

→ Proficiency Rating 진행

→ SET 5: 영어 문장 발화, 평서문 10개의 문장 선택

데이터 등급 분류 절차

→ 4명의 평가자

- 구글 설문 조사 (google survey)
 - Language Background
 - Basic information
- 각 파일 당 1~6 점으로 평가

→ 평가 파일

- 한국인의 영어 발화 2번 평가
- fillers 포함
- 랜덤화

해당 녹음 파일을 들은 뒤, 영어 문장 강세와 자연스러움에 대한 유창성을 1에서 6까지 나타내어 주세요.
(1-2 = beginner, 3-4 = intermediate, 5-6 = advanced). 가능한 고르게 분포해주시면 감사하겠습니다.
1 = very unnatural, 2 = unnatural, 3 = natural, 4 = fluent 5 = very fluent 6 = native-like

	1	2	3	4	5	6
파일 21	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
파일 22	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
파일 23	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
파일 24	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
파일 25	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
파일 26	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
파일 27	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
파일 28	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
파일 29	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
파일 30	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

데이터 등급 분류 절차

→ 등급 별 파일 구분


- 파일 간 평가 차이가 클 경우, 해당 평가는 분석에서 제외
- 평가들 간 평가 차이가 클 경우, 해당 파일은 분석에서 제외

→ 등급 별 스케일 및 파일 수

Group	Fluency Sacle	Files
Advanced	4.5 ~ 6	42
Intermediate	2.5 ~ 4.4	35
Novice	1 ~ 2.4	84


데이터 분석 절차

1) 강세 간 길이 측정

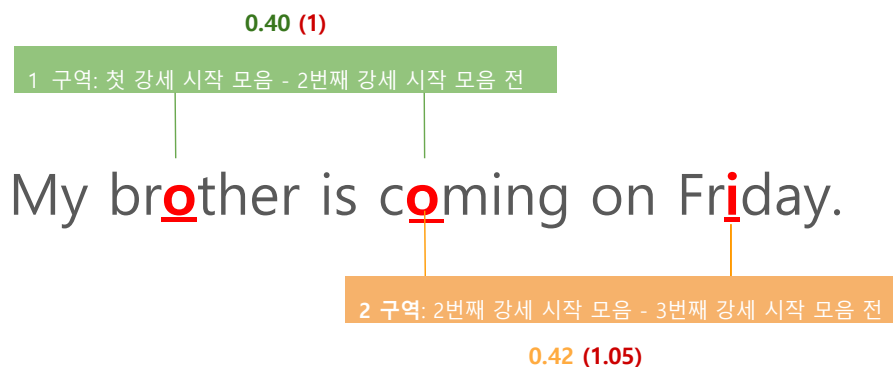
- 각 문장 내 강세 간 길이(duration) 측정
- 프로그램: Praat 

2) 길이의 비율(ratio) 계산

3) 표준 편차 (standard deviation) 통계 분석

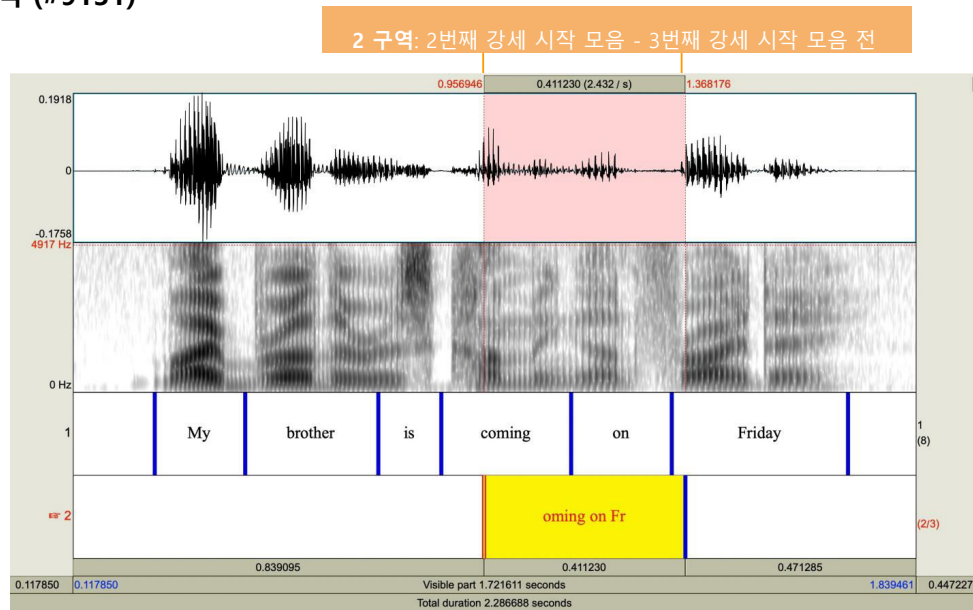
- 각 파일 당 표준 편차를 구한 뒤, 문장이 길어짐에 따라 isochrony의 구현 정도가 등급 별로 어떻게 달라지는 지를 보기 위해 **feet** 개수 별로 파일을 나누어 통계 분석을 진행
- 프로그램: one-way ANOVA, JASP 

데이터 분석 (#9131)



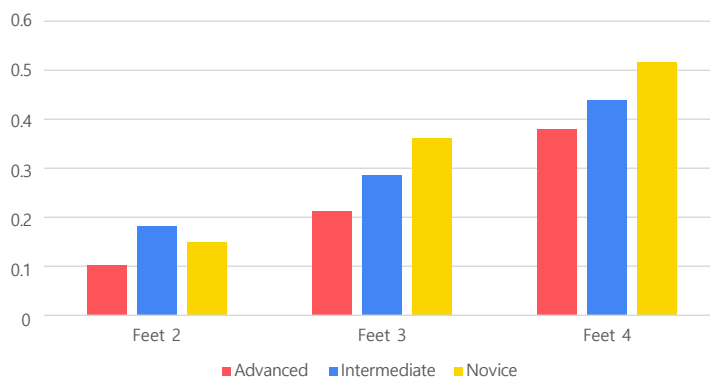
전체 (발화 시작-끝): 1.36초

데이터 분석 (#9131)



분석 결과

→ Descriptive Statistics 결과



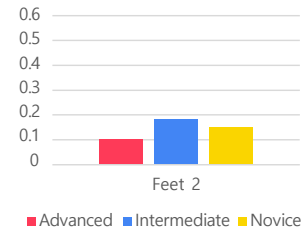
→ 모든 문장에서 상위(advanced) 그룹이 가장 0에 가깝게 발화

분석 결과

→ Sentences with 2 feet

Table 2. Results of descriptive statistics on isochrony for sentences with two feet

Group	N	Mean	SD	SE	Coefficient of variation
Advanced	22	0.103	0.097	0.021	0.947
Intermediate	18	0.182	0.150	0.035	0.826
Novice	42	0.149	0.104	0.016	0.700



→ 상 - 하 - 중 순서로 0에 가까움

Table 3. Results of one-way ANOVA on isochrony for sentences with two feet

Cases	Sum of Squares	df	Mean Square	F	p	η^2
Group	0.064	2	0.032	2.466	0.091	0.059
Residuals	1.028	79	0.013			

Note. Type III Sum of Squares

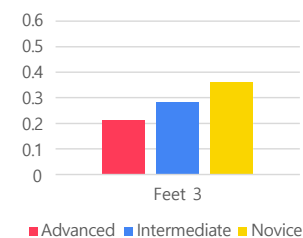
→ 그룹 간 유의미한 차이(significant difference)는 없음

분석 결과

→ Sentences with 3 feet

Table 4. Results of descriptive statistics on isochrony for sentences with three feet

Group	N	Mean	SD	SE	Coefficient of variation
Advanced	16	0.212	0.153	0.038	0.722
Intermediate	14	0.285	0.272	0.073	0.956
Novice	30	0.362	0.370	0.067	1.020



→ 상 - 중 - 하 순서로 0에 가까움

Table 5. Results of one-way ANOVA on isochrony for sentences with three feet

Cases	Sum of Squares	df	Mean Square	F	p	η^2
Group	0.244	2	0.122	1.315	0.276	0.044
Residuals	5.277	57	0.093			

Note. Type III Sum of Squares

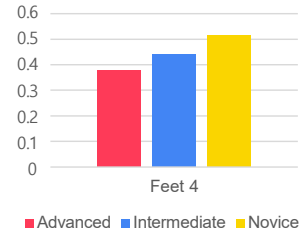
→ 그룹 간 유의미한 차이(significant difference)는 없음

분석 결과

→ Sentences with 4 feet

Table 6. Results of descriptive statistics on isochrony for sentences with four feet

Group	N	Mean	SD	SE	Coefficient of variation
Advanced	4	0.379	0.116	0.058	0.307
Intermediate	3	0.439	0.224	0.129	0.510
Novice	12	0.516	0.353	0.102	0.684



→ 상 - 중 - 하 순서로 0에 가까움

Table 7. Results of one-way ANOVA on isochrony for sentences with four feet

Cases	Sum of Squares	df	Mean Square	F	p	η^2
Group	0.061	2	0.031	0.324	0.728	0.039
Residuals	1.509	16	0.094			

Note. Type III Sum of Squares

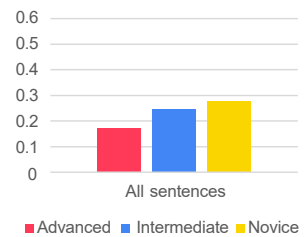
→ 그룹 간 유의미한 차이(significant difference)는 없음

분석 결과

→ 전체 문장 분석 결과

Table 8. Results of descriptive statistics on isochrony for all sentences

Group	N	Mean	SD	SE	Coefficient of variation
Advanced	42	0.171	0.147	0.023	0.865
Intermediate	35	0.245	0.221	0.037	0.900
Novice	84	0.278	0.298	0.033	1.073



→ 상 - 중 - 하 순서로 0에 가까움

Table 9. Results of one-way ANOVA on isochrony for all sentences

Cases	Sum of Squares	df	Mean Square	F	p	η^2
Group	0.322	2	0.161	2.564	0.080	0.031
Residuals	9.914	158	0.063			

Note. Type III Sum of Squares

→ 그룹 간 유의미한 차이(significant difference)는 없음

결론 및 한계

→ 결론

- 모든 문장에서 isochrony가 상-중-하 등급 간에 유의미한 차이를 보이지 않음
- 영어의 isochrony는 학습자의 능숙도 등급 구별에 큰 역할을 하지 않음

→ 한계

- 데이터의 문제
 - 데이터의 양(quantity)의 부족
 - 데이터의 질(quality)의 문제

→ 후속 연구를 통해 보다 정확한 결과를 얻고자 함

Reference

- Abercrombie, D. (1967). Elements of General Phonetics. Edinburgh University Press.
- Chen, Y., Robb, M. P., Gilbert, H. R., & Lerman, J. W. (2001). A study of sentence stress production in Mandarin speakers of American English. *The Journal of the Acoustical Society of America*, 109(4), 1681-1690.
- Choi, H. S. (2011). Acoustic Measurement of English read speech by native and nonnative speakers. *Phonetics and Speech Sciences*, 3(3), 77-88.
- Chun, D. M. 2002. Discourse Intonation in L2: From Theory and Research to Practice. Amsterdam: John Benjamins Publishing Company.
- JASP Team. (2021). JASP (Version 0.15) [Computer software]
- Kim, O. Y. (2007). An acoustic study of English sentence stress and rhythm produced by Korean speakers. *Speech Sciences*, 14(1), 121-135.
- Kwon, H. R. (2007). A phonetic analysis of advanced Korean EFL learners' English reduced vowels. *Korean Journal of English Language and Linguistics*, 7(4), 547-564.
- Ladefoged, P., & Johnson, K. (2011). *A Course in Phonetics*, 6th Edition, pp.38-44
- Lee, J. P., & Jang, T. Y. (2004). A comparative study on the production of inter-stress intervals of English speech by English native speakers and Korean speakers. In *Eighth International Conference on Spoken Language Processing*.
- Lee, H. Y., & Song, J. (2019). Evaluating Korean learners' English rhythm proficiency with measures of sentence stress. *Applied Psycholinguistics*, 40(6), 1363-1376.
- Mochizuki-Sudo, M., & Kiritani, S. (1991). Production and perception of stress-related durational patterns in Japanese learners of English. *Journal of Phonetics*, 19(2), 231-248.
- Park, H. L. (2021). Stress and vowel reduction by Korean Learners of English (Doctoral dissertation, Colorado State University).
- Pike, K. L. (1946). *The Intonation of American English*. Ann Arbor: University of Michigan Press.
- Rhee, S. C., Lee, S. H., Kang, S. K., & Lee, Y. J. (2003). Design and construction of Korean-Spoken English Corpus (K-SEC). *Malsori*, 46, 159-174.
- Ueyama, M. (2000). *Prosodic transfer: An acoustic study of L2 English vs. L2 Japanese*. University of California, Los Angeles.
- Woo, S. (2020). The Recognition of English Sentence Stress by Elementary School Pre-Service Teachers. *The Journal of Studies in Language*, 35(4), 503-516.
- Yang, B. G. (2002). An acoustical study of English word stress produced by Americans and Koreans. *Speech Sciences*, 9(1), 77-88.
- Yang, Inyoung (2016). Korean learners' production of English sentence stress: A syntagmatic observation. *Journal of the Korea English Education Society*, 15(1), 53-79.
- Yoo, H. B. (2018). An Acoustic Analysis of English Prosody by Mongolian Learners of English. *English Language and Linguistics*, 30, 153-177.
- Yoo, H. B. (2014). Phonetic reduction of English function words in the passage reading by Korean EFL learners. *English Language and Linguistics*, 20(1), 179-202.
- Yun, G. (2020). English Native and Nonnative Speakers' Perception of English Reduced Word Forms with Reduced Vowels. *Journal of Asia TEFL*, 17(1), 194.

Thank you 😊



Examining the Transformer Models with the K-CoLA Dataset

Yong-hun Lee

*Department of Linguistics
Chungnam National University*

Yebin Lee

*Department of Linguistics
Chungnam National University*

Examining the Transformer Models with the K-CoLA Dataset

1. Introduction

■ Motivation

- ◆ The Transformer models are used in many downstream tasks including NLP.
- ◆ Since the CoLA (Corpus of Linguistic Acceptability) dataset was included in the GLUE Benchmark (Wang et al., 2019), there have been several trials to apply the Transformer models in the acceptability tests.

◆ Questions

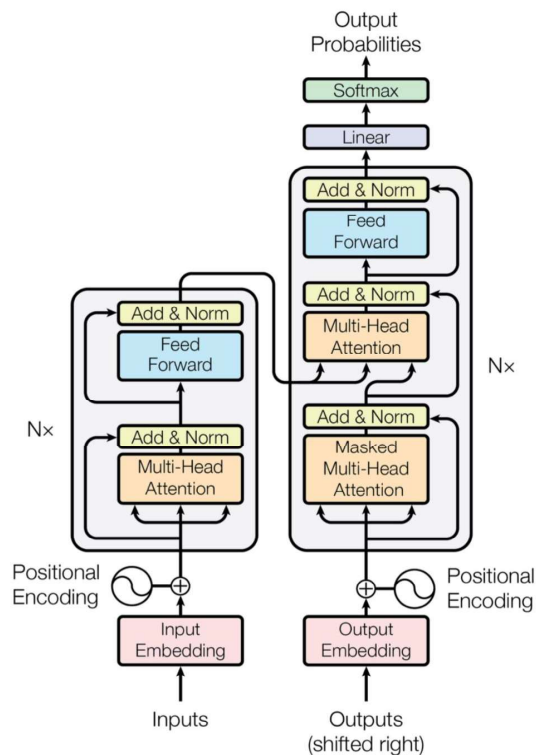
- Is it possible to apply the techniques of analyzing the CoLA dataset to the Korean data?
- Are the performance of three Transformer models (mBERT, KoBERT, and KR-BERT) similar, especially to the specific type of phenomenon?

■ Goal

- ◆ To examine the performance of three types of Transformer models (mBERT, KoBERT, and KR-BERT) for various types of linguistic phenomena with the K-CoLA dataset.

2. Previous Studies

■ Transformer Model (Vaswani et al., 2017)



2. Previous Studies

■ Three Types of BERT Models for the Korean Language

- ◆ mBERT
 - Multilingual BERT
 - Pires et al. (2019)
 - Pretrained model on the top 104 languages with the largest Wikipedia using a masked language modeling (MLM) objective
- ◆ KoBERT
 - Korean BERT
 - SK Telecom (2019)
 - Developed to overcome the Korean performance limitation of existing BERT
 - Learned a large-scale corpus of millions of Korean sentences collected from Wikipedia or news, and applied a data-based tokenization technique to reflect the characteristics of irregular language changes in Korean, using only 27% of tokens compared to the previous one
 - Led to a performance improvement of more than 2.6%
- ◆ KR-BERT
 - KoRean-based BERT pre-trained
 - Lee et al. (2020)
 - Korean-specific, small-scale, with comparable or better performances developed by Computational Linguistics Lab at Seoul National University

2. Previous Studies

■ The CoLA Dataset

- ◆ The Corpus of Linguistic Acceptability
- ◆ Warstadt et al. (2018)
- ◆ It is included in the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2019).
- ◆ The dataset consists of 10,657 sentences from 23 linguistics publications, expertly annotated for acceptability (grammaticality) by their original authors.
- ◆ Warstadt and Bowman (2020) annotated the CoLA dataset with major and minor features.

■ The K-CoLA Dataset

- ◆ The Korean version of the CoLA dataset
- ◆ Ahn (2020)
- ◆ The dataset consists of 18,099 sentences.

2. Previous Studies

■ A Random Sample of Sentences (CoLA)

Acceptability Sentence	Simple	Locative	PP Arg-VP	High Arity	Passive	Binding:Other	Emb Q	Complex QP	Modal	Raising	Trans-Adj	Coord	Ellipsis/Anaphor	Comparative
✓ The magazines were sent by Mary to herself.			X	X	X									
✓ John can kick the ball.									X					
* I know that Meg's attracted to Harry, but they don't know who.			X				X					X	X	
✓ They kicked them	X					X								
✓ Which topic did you choose without getting his approval?								X						
* It was believed to be illegal by them to do that.				X	X					X	X			
* Us love they.	X													
* The more does Bill smoke, the more Susan hates him.						X								X
✓ I ate a salad that was filled with lima beans.			X		X									
✓ That surprised me.	X													

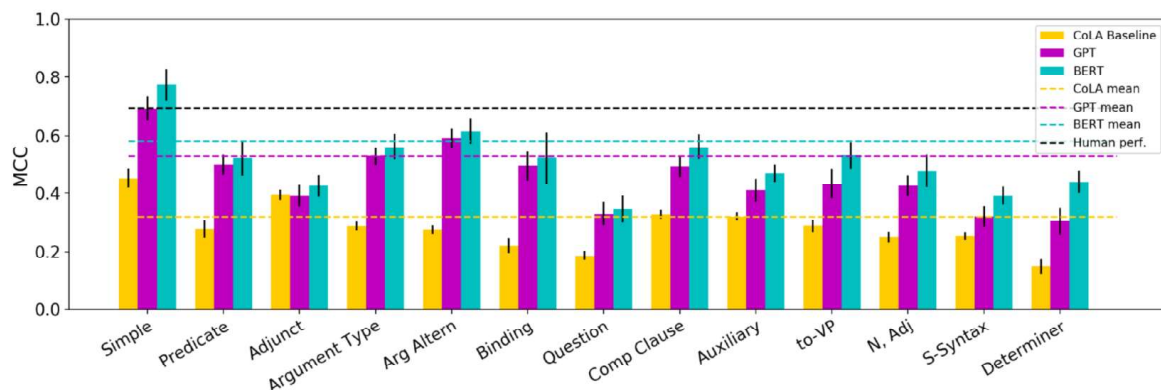
2. Previous Studies

Major and Minor Features (CoLA)

Major Feature (<i>n</i>)	Minor Features (<i>n</i>)
Simple (87)	Simple (87)
Pred (256)	Copula (187), Pred/SC (45), Result/Depictive (26)
Adjunct (226)	VP Adjunct (162), Misc Adjunct (75), Locative (69), NP Adjunct (52), Temporal (49), Particle (33)
Arg Types (428)	PP Arg VP (242), Oblique (141), PP Arg NP/AP (81), Expletive (78), by-Phrase (58)
Arg Altern (421)	High Arity (253), Passive (114), Drop Arg (112), Add Arg (91)
Bind (121)	Binding:Other (62), Binding:Refl (60)
Question (222)	Emb Q (99), Pied Piping (80), Rel Clause (76), Matrix Q (56), Island (22)
Comp Clause (190)	CP Arg VP (110), No C-izer (41), Deep Embed (30), CP Arg NP/AP (26), Non-finite CP (24), CP Subj (15)
Auxiliary (340)	Aux (201), Modal (134), Neg (111), Psuedo-Aux (26)
to-VP (170)	Control (80), Non-finite VP Misc (38), VP Arg NP/AP (33), VP+Extract (26), Raising (19)
N, Adj (278)	Comp NP (106), Rel NP (65), Deverbal (53), Trans Adj (39), NNCompd (35), Rel Adj (26), Trans NP (21)
S-Syntax (286)	Coord (158), Ellipsis/Anaphor (118), Dislocation (56), Subordinate/Cond (41), Info Struc (31), S-Adjunct (30), Frag/Paren (9)
Determiner (178)	Quantifier (139), NPI/FCI (29), Comparative (25), Partitive (18)

2. Previous Studies

Performance (Matthews Correlation Coefficient; MCC)



3. Research Method

■ Dataset

- ◆ Training Dataset: K-CoLA (Ahn, 2020)
- ◆ Test/Target Dataset: Korean translation of the dataset in Warstadt and Bowman (2020)
- ◆ Acceptability (Human, mBERT, KoBERT, and KR-BERT)
- ◆ Only the major features in Warstadt and Bowman (2020)

■ Models

- ◆ mBERT
- ◆ KoBERT
- ◆ KR-BERT

3. Research Method

■ Procedure

- ◆ Download the deep learning models
 - mBERT: <https://github.com/google-research/bert/blob/master/multilingual.md>
 - KoBERT: <https://github.com/SKTBrain/KoBERT>
 - KR-BERT: <https://github.com/snunlp/KR-BERT>
- ◆ Prepare the training dataset (Ahn, 2020) and the test dataset (Korean Translation of Warstadt and Bowman (2020))
- ◆ Train the BERT models with the training dataset (10 epochs, LR=1e-5)
- ◆ Measure the acceptability label (0 or 1) for each sentence in the test dataset (for all of the three models)
- ◆ Evaluate the models with the following metrics
 - Precision, Recall, F1, and Accuracy
 - Matthews Correlation Coefficient (MCC)

3. Research Method

■ Human Annotation for Golden Standard

- ◆ English vs. Korean (4 Possibilities)
 - English acceptable and Korean unacceptable: None
 - English acceptable and Korean acceptable: Most
 - English unacceptable and Korean unacceptable
 - ▶ *I ate that she was happy.
 - ▶ *나는 그녀가 행복하다고 먹었다.
 - English unacceptable and Korean acceptable
 - ▶ *The bird sing.
 - ▶ 새가 노래합니다.
- ◆ Human Annotation
 - 2 Native Speakers
 - Non-linguistic Major
 - Acceptable only when both annotators marked the sentence acceptable

4. Analysis Results

■ Confusion Matrix

	Predicted Positives	Predicted Negatives
Positives	True Positives	False Negatives
Negatives	False Positives	True Negatives

4. Analysis Results

■ Precision, Recall, F1, and Accuracy

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\text{-score} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

■ Matthews Correlation Coefficient (MCC)

$$MCC = \frac{TN \times TP - FN \times FP}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

4. Analysis Results

■ mBERT

Model	Phenomenon	Sentences	Precision	Recall	F1	Accuracy	MCC
mBERT	All	1043	0.153	0.241	0.188	0.826	0.098
mBERT	Simple	87	0.182	0.333	0.235	0.851	0.169
mBERT	Predicate	256	0.132	0.227	0.167	0.805	0.068
mBERT	Adjunct	226	0.133	0.235	0.170	0.827	0.086
mBERT	ArgumentType	428	0.161	0.231	0.189	0.820	0.094
mBERT	ArgAltern	421	0.134	0.273	0.180	0.805	0.091
mBERT	Imperative	12	0.000	NaN	NaN	0.750	NaN
mBERT	Binding	121	0.143	0.400	0.211	0.752	0.120
mBERT	Question	222	0.400	0.308	0.348	0.797	0.233
mBERT	CompClause	190	0.091	0.200	0.125	0.779	0.020
mBERT	Auxiliary	340	0.195	0.222	0.208	0.821	0.107
mBERT	to_VP	170	0.143	0.214	0.171	0.829	0.083
mBERT	N_Adj	278	0.235	0.286	0.258	0.835	0.167
mBERT	S_Syntax	286	0.261	0.308	0.282	0.787	0.159
mBERT	Determiner	178	0.143	0.125	0.133	0.854	0.054
mBERT	Violations	145	0.113	0.400	0.176	0.614	0.024

4. Analysis Results

KoBERT

Model	Phenomenon	Sentences	Precision	Recall	F1	Accuracy	MCC
KoBERT	All	1043	0.007	0.250	0.014	0.867	0.022
KoBERT	Simple	87	0.000	NaN	NaN	0.874	NaN
KoBERT	Predicate	256	0.000	0.000	NaN	0.848	-0.026
KoBERT	Adjunct	226	0.033	0.500	0.062	0.867	0.102
KoBERT	ArgumentType	428	0.000	0.000	NaN	0.864	-0.027
KoBERT	ArgAltern	421	0.015	0.500	0.029	0.841	0.064
KoBERT	Imperative	12	0.000	NaN	NaN	0.750	NaN
KoBERT	Binding	121	0.000	NaN	NaN	0.769	NaN
KoBERT	Question	222	0.000	0.000	NaN	0.860	-0.027
KoBERT	CompClause	190	0.000	NaN	NaN	0.826	NaN
KoBERT	Auxiliary	340	0.000	0.000	NaN	0.874	-0.028
KoBERT	to_VP	170	0.000	0.000	NaN	0.871	-0.029
KoBERT	N_Adj	278	0.029	0.333	0.054	0.874	0.067
KoBERT	S_Syntax	286	0.000	0.000	NaN	0.832	-0.037
KoBERT	Determiner	178	0.000	NaN	NaN	0.921	NaN
KoBERT	Violations	145	0.000	0.000	NaN	0.628	-0.063

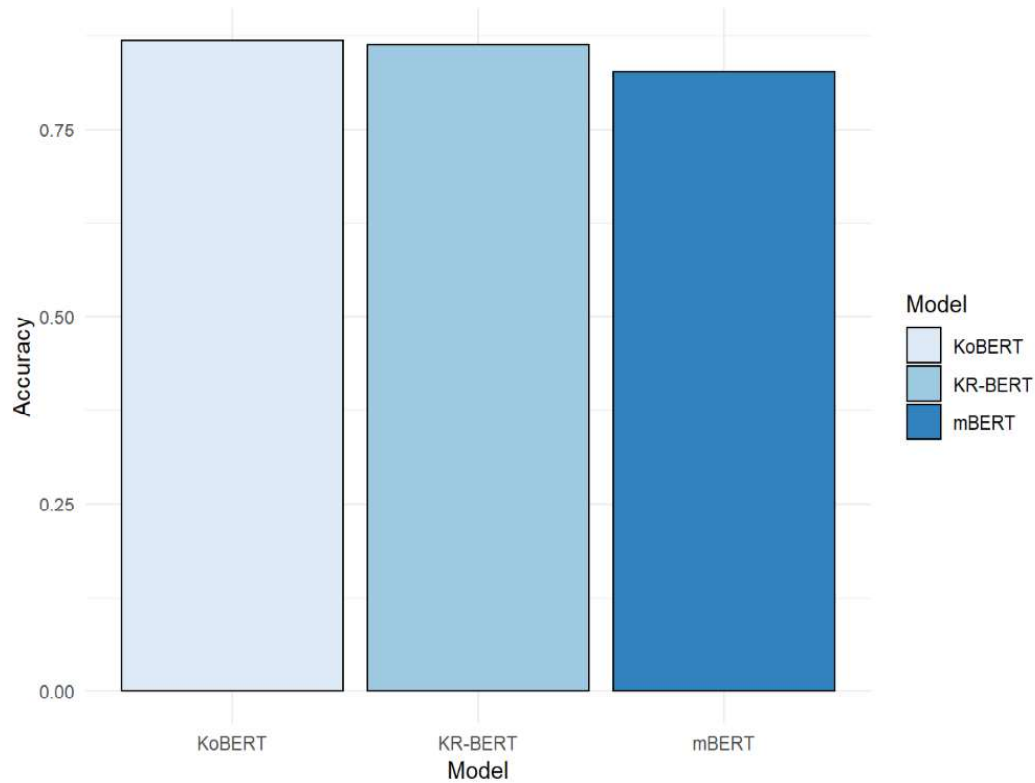
4. Analysis Results

KR-BERT

Model	Phenomenon	Sentences	Precision	Recall	F1	Accuracy	MCC
KR-BERT	All	1043	0.015	0.182	0.027	0.862	0.015
KR-BERT	Simple	87	0.000	NaN	NaN	0.874	NaN
KR-BERT	Predicate	256	0.000	0.000	NaN	0.848	-0.026
KR-BERT	Adjunct	226	0.033	0.333	0.061	0.863	0.069
KR-BERT	ArgumentType	428	0.018	0.167	0.032	0.860	0.013
KR-BERT	ArgAltern	421	0.030	0.400	0.056	0.838	0.072
KR-BERT	Imperative	12	0.000	0.000	NaN	0.667	-0.174
KR-BERT	Binding	121	0.000	0.000	NaN	0.752	-0.071
KR-BERT	Question	222	0.000	0.000	NaN	0.851	-0.046
KR-BERT	CompClause	190	0.000	0.000	NaN	0.821	-0.033
KR-BERT	Auxiliary	340	0.024	0.167	0.043	0.868	0.019
KR-BERT	to_VP	170	0.000	0.000	NaN	0.871	-0.029
KR-BERT	N_Adj	278	0.029	0.333	0.054	0.874	0.067
KR-BERT	S_Syntax	286	0.000	0.000	NaN	0.825	-0.052
KR-BERT	Determiner	178	0.071	0.333	0.118	0.916	0.124
KR-BERT	Violations	145	0.000	0.000	NaN	0.614	-0.110

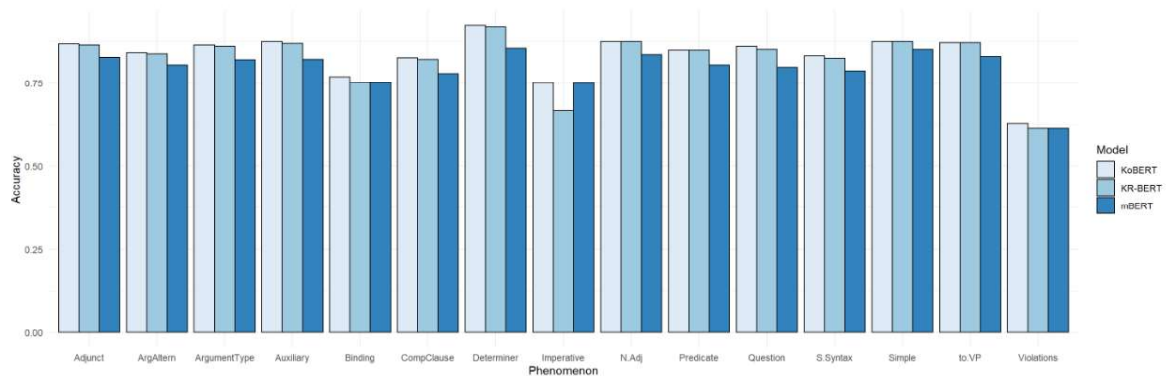
4. Analysis Results

Overall Performance



4. Analysis Results

Performance for Individual Phenomena



5. Conclusion

Summary

- ◆ We trained three types of Transformer models (mBERT, KoBERT, and KR-BERT) with the K-CoLA dataset (Ahn et al., 2020).
- ◆ We used the Korean translation of the annotated CoLA dataset (Warstadt and Bowman, 2020) and measured the performance of the three types of BERT models.
- ◆ Overall Performance
 - KoBERT > KR-BERT > mBERT
- ◆ Performance for Specific Phenomenon
 - KoBERT > KR-BERT > mBERT (in general)
 - The worst performance in 'Violations'

Questions

- ◆ Korean-specific phenomena?
- ◆ Necessary to use the same criteria?
- ◆ Necessary to develop a new set of criteria?

References

- Ahn, H. 2020. *Construction of the Data for the Acceptability Judgment of Korean Sentences*. Seoul: National Institute of Korean Language.
- Lee, S., H. Jang, Y. Baik, S. Park, and H. Shin. 2020. KR-BERT: A small-scale Korean-specific language model. arXiv preprint arXiv:2008.03979.
- Pires, T., E. Schlinger, and D. Garrette. 2019. How multilingual is multilingual BERT? arXiv preprint arXiv:1906.01502.
- Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention Is All You Need. arXiv preprint arXiv:1706.03762.
- Wang, A., A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. 2019. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. arXiv preprint arXiv:1804.07461.
- Warstadt, A., A. Singh, and S. Bowman. 2018. Neural Network Acceptability Judgments. arXiv preprint arXiv:1805.12471.
- Warstadt, A. and S. Bowman. 2020. Linguistic Analysis of Pretrained Sentence Encoders with Acceptability Judgments. arXiv preprint arXiv: 1901.03438.

Meeting between artificial intelligence and corpus linguistics: Collocation analysis for method validation of AI

김원빈 (연세대학교 영어영문학과)

The purpose of this study is to diagnose the current level of artificial intelligence (AI) techniques and propose collocation analysis in corpus linguistics (CL) as a methodology for the method validation of AI-based analysis. For this purpose, this study tracks semantic change of the two Korean neologisms *leyal* and *lwuce* using machine learning techniques and compares their results with those from collocation analysis. To analyze the semantic changes of the two neologisms, I collected Korean Twitter data and constructed ten yearly Twitter corpora from 2010 to 2019 for each neologism. For *leyal*, I used word2vec and cosine similarity to measure (i) semantic similarity between *leyal* and *cincca*, an alternative Korean word representing one of the two meanings of *leyal* and (ii) semantic similarity between *leyal* and *leyalmatulitu*, an alternative Korean word representing the other meaning of *leyal* for each of the ten years. The results showed that the meaning of the neologism *leyal* is more similar to that of *cincca*, which implies that *leyal* has been more used to represent ‘real’ than ‘Real Madrid’. Also, they showed that there is no semantic change in terms of the relation between the two meanings (‘real’ and ‘Real Madrid’). These results agreed with those from the analysis of top collocates of *leyal*. For *lwuce*, I employed latent dirichlet allocation to measure the number of topics in each of the two subcorpora classified depending on two meanings of *lwuce* (i.e., ‘a person who is defeated or unsuccessful’ and ‘a man whose stature is below 180 centimeters’) for each year. The results indicated that the use of former meaning sharply increased in 2015 and the use of latter meaning increased in 2018. In addition, there was no semantic change concerning the relation between the two meanings. However, the analysis of top collocates of *lwuce* did not demonstrate the increased use of latter meaning in 2018. The partial agreement of results means that the current level of computer analysis still falls short of human analysis. Also, it implies that we should employ AI techniques carefully and accept their results prudently. This study is the first work attempting to establish a specific method to evaluate the suitability of AI approach for semantic change research using collocation analysis in CL. This attempt is expected to make a contribution to bridging a gap between NLP and CL.

Keywords: artificial intelligence, collocation analysis, method validation, word2vec, latent dirichlet allocation

1. Introduction

Artificial intelligence (AI) is one of the hottest issues today. As advanced technology performs laborious tasks for humans, we can save much time and energy. This advantage applies in natural language processing (NLP) as well. Analysis of large-scale corpora using AI techniques is faster, more convenient, and more efficient, compared to a manual analysis. However, such analysis has lower accuracy than human analysis because computers use not “language” but “numbers” to process human language. In the process of converting “language” into “numbers” and processing them, computers rely on statistics without taking into account any linguistic features, language exceptions and variables. Thus, such analysis is not the same as analysis by linguists. Considering this imperfection of AI techniques, this study aims to diagnose the current level of computer analysis by comparing results between AI technique-based analysis and existing manual linguistic analysis (i.e., collocation analysis). Also, this study suggests collocation analysis as a new methodology for method validation of the AI approach. As of now, methods associated

with corpus linguistics have never been tried for method validation of NLP techniques¹. Therefore, the application of collocation analysis is expected to play a significant role in connecting corpus linguistics with NLP. For those purposes, this study employs two machine learning techniques, i.e., word2vec and latent dirichlet allocation to track semantic change of the two Korean neologisms *leyal*² (레알) and *lwuce* (루저).

2. Analysis of semantic change of *leyal*

2.1 Data collection and preprocessing

I employed snsrape in Python to collect 15,500 Korean tweets including *leyal* (레알) as a keyword for each month over the course of ten years, from 2010 to 2019. A total of 186,000 tweets was scraped for each year except for four years. The numbers of tweets scraped for the four atypical years (i.e., 2010, 2015, 2016, and 2019) fell below the target amount. Based on the scraped tweets, I built ten yearly Twitter corpora. For data preprocessing, I removed unnecessary data, tokenized sentences into morphemes, reduced inflected morphemes down to their stems, and got rid of stop words. The process of data preprocessing was performed by means of KoNLPy.

2.2 Methodology

The neologism *leyal* appeared in late 2009, which came from the English word *real*. *Leyal* is used to mean ‘really’ but depending on the contexts, it refers to the Spanish professional football club Real Madrid. To find out how the two meanings of *leyal* have developed across time and which meaning is more dominant, I used two alternative Korean words, *cincca* (진짜) representing ‘really’ and *leyalmatulitu* (레알마드리드) representing ‘Real Madrid’. On the basis of the statement from Harris (1954) that the contexts of semantically similar words are similar to each other, I compared the context words of *leyal* and those of *cincca* and *leyalmatulitu*. If the context of *leyal* is more similar to that of *cincca*, it can be said that *leyal* is more semantically similar to *cincca*, implying that the meaning ‘really’ is more dominant than the meaning ‘Real Madrid’. However, it is inefficient to manually compare the context words among the three words so I utilized vectors from word2vec. Word2vec converts words into vectors, considering context words of the target word. Thus, the comparison between vectors means comparison between context words. After training the word2vec model with each of the ten Twitter corpora, I obtained vectors of the three words. Then, I investigated cosine similarity (CS) values between the vectors of *leyal* and *cincca* and CS values between the vectors of *leyal* and *leyalmatulitu*. This is because CS measures similarity between vector (i.e., semantic similarity between words). The closer the CS value is to “1,” the more similar the meanings of the two words are. To track the semantic change of *leyal*, I observed the change of CS values over the ten years.

2.3 Results

Figure 1 shows the change of CS values across time. The CS values between *leyal* and *cincca* are consistently closer to “1” over the ten years. However, none of the CS values between *leyal* and *leyalmatulitu* exceed “0.306”. This indicates that *leyal* has a greater similarity in meaning to *cincca*, suggesting that *leyal* has been used more frequently to convey the meaning of ‘really’

¹ This study is a part of my dissertation (Kim 2022). Before my dissertation research, methods related to corpus linguistics had not been utilized for the method validation of NLP techniques. Additionally, this study was partly presented at the 2023 Winter Conference of the Korean Association of Language Sciences (Kim 2023).

² Korean words in this paper have been romanized conforming to the Yale system of romanization.

rather than ‘Real Madrid’. While the CS values between *leyal* and *cincca* exhibit relative stability, the CS values between *leyal* and *leyalmatulitu* show more frequent and significant fluctuations. This suggests that the use of *leyal* to convey the meaning of ‘really’ has remained consistent, whereas the use of *leyal* to refer to Real Madrid is comparatively unstable and has not become firmly established. The variation in the CS values over time indicates that each meaning of *leyal* undergoes subtle changes from year to year. However, concerning the relationship between the two meanings, there has been no semantic change. The meaning of ‘really’ has consistently been more dominant than that of ‘Real Madrid’ throughout the ten-year period.

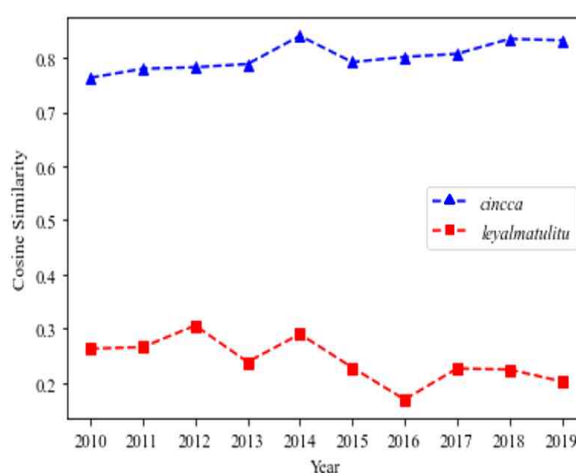


Figure 1 Change in the cosine similarity values between *leyal* and *cincca* and those between *leyal* and *leyalmatulitu* across time

2.4 Method validation

For method validation, I used collocation analysis and a notion of A-curve in language as a complex system. To be specific, I obtained collocates of *leyal* for each corpus by means of LancesBox. Next, I arranged them according to their frequencies. As Kretzschmar argues, their frequency profiles showed an “A-curve”. According to Kretzschmar (2009, 2015), the frequency distribution of variants for any given linguistic feature consistently follows a nonlinear pattern, i.e., an “A-curve” (Figure 2). The A-curve consists of a small number of variants with high frequencies, a moderate number of variants with moderate frequencies, and the majority of variants with very low frequencies. Because a small number of variants with high frequencies are essential variants characterizing the entire variants for the given linguistic feature, their change suggests a significant change to that linguistic feature.

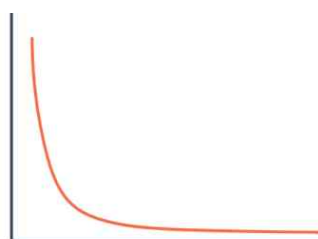


Figure 2 A-curve

Based on the notion of A-curve, I extracted the top thirty collocates of *leyal* from each corpus and compared them. For the systematic observation of change to their order, I investigated how the order of the top thirty collocates from the 2010 corpus changes over time. Table 1 indicates that nineteen out of the top twenty collocates from the 2010 corpus are consistently used with high frequency in the other corpora and the change in the order of the collocates is not significant. Also, it shows that frequent collocates continue to be used with high frequency, while infrequent collocates continue to be used infrequently, which suggests that there has been no significant change in the way people use *leyal*. The findings demonstrate that there has been no notable semantic change in *leyal* over the ten years as well as the usage of *leyal* has remained consistent.

English meaning	2010 Collocate	Frequency	2011	2012	2013	2014	2015	2016	2017	2018	2019
Really	레알	176831	179274	172432	160215	163407	166130	162932	174787	174533	168342
To do	하다	60943	54655	54309	53356	58493	59492	60529	67041	73390	75741
To see/watch/look	보다	21376	21744	20452	19343	21277	21312	20623	21443	23072	24363
To exist/have	있다	16123	14744	13696	13937	15338	15134	14582	14765	16423	18098
Really	진짜	13555	19427	20757	19457	20703	23744	25920	34601	38240	37916
To become/be	되다	11880	10866	10300	9924	10705	10373	10155	10305	11224	12014
Not to be	아니다	11212	10836	10960	10531	11817	12861	14221	16192	19441	21023
All	다	10259	11116	11606	10887	11774	12179	13487	16485	17604	17532
To be good	좋다	10189	10124	9716	9398	10259	10974	10148	10222	10940	10613
Not to exist/have	없다	9732	9947	9647	9838	11221	10978	11053	11676	13105	13335
-	님	9458	7126	6860	7303	7547	8310	7908	8922	9426	9342
To be so	그렇다	9361	9076	8043	7720	8197	7994	7686	7477	8277	8512
Not	안	9341	9810	9566	9546	10785	11170	11487	12924	14057	14402
To go	가다	9338	9387	8094	7230	6770	6913	6953	7948	8156	8167
My	내	9186	10457	10792	9772	10394	10286	10833	11704	11595	11246
One	한	9099	8432	8042	7982	8717	8342	8461	8749	9739	10451
-	임	8671	7730	6674	5992	5546	5159	5934	8242	8839	8727
Today	오늘	8609	7708	6637	5264	5046	5177				
To come	오다	8112	7684	7072	7127	7788	8738	9174	12272	12484	12067
Thing	거	8036	7558	8407	9062	13332	19989	21650	21503	19977	20375
I	난	7343	6536	5370							
-	당	7073									
By the way/But	근데	7042	8765	8595	8156	7871	7648	7063	7496	8293	9185
To eat	먹다	6973	6405		4787	5351	5495			6471	7062
-	이야	6591	7145	6368	5059						
Really/Truly	정말	6582									
To break out	뜯다	6370									
Cannot	못	6204	5996	5940	5701	6550	6868	7254	8077	8827	8710
-	를다	6199	5809	5281	5093	5265	5473	5819	6381	7310	7397
To sleep	자다	5925									

Table 1 Top thirty collocates of *leyal* from the 2010 corpus and change in their frequencies across time

In order to find out the dominance relation of the two meanings ‘really’ and ‘Real Madrid’, I compared the top thirty collocates of *leyal* from the 2010 corpus with those of *cincca* and *leyalmatulitu* from the same corpus (due to the page limit, I omit the tables related to the comparison here). Interestingly, the top thirty collocates of *leyal* were the same as those of *cincca*. This means that their meanings are similar to each other and *leyal* has been used to represent ‘really’. Since Table 1 showed that there has been no semantic change over the ten years, this finding from the 2010 corpus will be able to apply in the other corpora as well. In contrast, some of the top thirty collocates of *leyalmatulitu* were different from those of *leyal*. This indicates that the usage of *leyal* differs from that of *leyalmatulitu* and their meanings are quite different. In conclusion, the findings from collocation analysis agree with those from word2vec and cosine similarity. Thus, we can say that the word2vec model has been well trained and its results are reliable.

2.5 Discussion

The findings from word2vec and cosine similarity demonstrate that there has been no substantial variation in the usage of the neologism *leyal*. This suggests that the utilization of *leyal* to mean ‘really’ is commonly accepted and considered normal. The agreement between the results from the AI approach and collocation analysis raises the question of why we should employ word2vec and cosine similarity (CS) despite similar outcomes. The primary advantage of employing word2vec is that it can take word meaning into account. Vectors generated by word2vec reflect word meaning. Based on these vectors, CS quantitatively provides information on the degree of similarity between the meanings of words. The quantification allows us to observe language change through numbers so we can analyze language change more clearly. Also, CS values make it more accessible and convenient to explore word meanings and conduct research in lexical semantics. The utilization of CS values between the neologism *leyal* and the two alternative words has significantly improved the efficiency of tracking the semantic change of *leyal*. This efficiency becomes particularly advantageous when analysts need to analyze vast amounts of data. Therefore, when analyzing large datasets that cover extended time periods, it is highly recommended to employ word2vec and cosine similarity techniques.

3. Analysis of semantic change of *lwuce*

3.1 Data collection and preprocessing

I used snsrape in Python to scrape tweets including *lwuce* (루저) as a keyword from 2010 to 2019. Because I could not collect the same number of tweets per year, I scraped tweets within a range of a minimum of 16,655 and a maximum of 27,387, with the number of monthly scraped tweets uneven. On the basis of the scraped tweets, I constructed ten yearly Twitter corpora. Next, I split each of the ten yearly Twitter corpora into two subcorpora depending on the two meanings of *lwuce* (i.e., ‘a person who is defeated or unsuccessful’ and ‘a man whose stature is below 180 centimeters’): the existing meaning subcorpus and the new meaning subcorpus, resulting in a total of twenty corpora. A tweet containing any of the following seven morphemes was classified as the new meaning subcorpus: 180, *khi* (키), *iha* (이하), *kkalchang* (갈창), *seynthi* (센티), *seynchi* (센치), and *sseynthi* (쎄티). For preprocessing, I removed unnecessary data, tokenized sentences into morphemes, reduced inflected morphemes down to their stems, and tagged them with their parts of speech. Using the tag information, I only extracted numbers and nouns which are not stop words from the corpora. This is because the performance of implementing latent dirichlet allocation (LDA) only with nouns is better. The process of data preprocessing was performed by means of KoNLPy.

3.2 Methodology

The word *lwuce* derived from the English word *loser* and its meaning is also similar to that of *loser*. It represents ‘a person who is defeated or unsuccessful’. However, in a Korean TV show aired in November 2009, a woman said that a man whose stature is below 180 centimeters is a loser. Back then, her remark caused a heated controversy and she was denounced for her statement. Since her remark, *lwuce* has been employed to refer to a man whose stature is below 180 centimeters. The neologism *lwuce* is a case where a new meaning is combined with the existing meaning. To find out how the two meanings of *lwuce* have developed across time and which meaning holds a predominant position, I used numbers of topics. As a specific word is frequently used, the number of tweets including the word will also increase. Many tweets contain relatively more contexts of the word than a smaller number of tweets. This leads to an increase

in the number of topics. Based on this, I assumed the variation in the number of topics will be able to show change in the use of word. To obtain the number of topics from each of the twenty preprocessed subcorpora, I employed C_v topic coherence. Because it is difficult to verify whether LDA has correctly classified lexical items based on topics, an evaluation metric to assess the performance of LDA is needed. If we blindly rely on the classification results from LDA without verification, we may obtain an incorrect number of topics from corpora with inaccurate topic classification. This is why I used C_v topic coherence to determine the number of topics. To be specific, I set the limit of trials for each subcorpus to fifty and selected the first number that met the following requirement on topic coherence within the fifty trials: The value of C_v topic coherence has to exceed the minimum value of .55, which is considered as a satisfactory value. On the basis of twenty number of topics from each subcorpus, I tracked how the two meanings of *lwuce* have developed and which meaning is more dominant.

3.3 Results

Figure 3 shows change in the proportions of topic numbers over the course of ten years. The proportions from the existing meaning subcorpora consistently surpass those from the new meaning subcorpora. This suggests that there has been no semantic change concerning the dominance relation between the two meanings. To be specific, the proportion from the new meaning subcorpus gradually decreased from 2010 to 2017 (with a slight increase in 2014) but significantly increased in 2018 and remained the same in 2019. This implies that there was a gradual decline in the use of *lwuce* to describe ‘a man whose stature is below 180 centimeters’ but it started to be used again in 2018. In addition, the difference in the proportions between 2014 and 2015 is greatest, which reveals that the most significant shift in the use of the two meanings occurred in 2015. The second most notable change occurred in 2018. The year 2018 witnessed an increase in the use of the new meaning and a decrease in the use of the existing meaning.

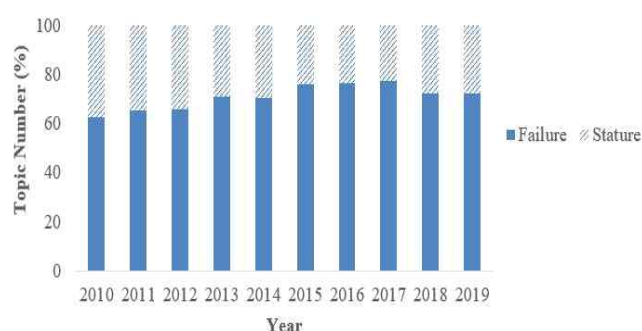


Figure 3 Change in the proportions of topic numbers across time

3.4 Method validation

I used LancsBox to obtain the collocates of *lwuce* from each of the ten preprocessed yearly corpora, with its two subcorpora (i.e., existing and new meaning subcorpora) combined for each corpus. Next, I arranged them according to their frequencies (their frequency profiles also followed A-curves). After selecting top thirty collocates from each corpus, I observed how the orders of three collocates, the number 180, the two collocates *namca* (남자), and *khi* (키), have

changed across the ten corpora. Table 2 shows that the collocate *khi* consistently holds the third position in ranking for four consecutive years from 2010 to 2014. Since 2015, its order has been fluctuating, moving beyond the tenth rank. The other two collocates 180 and *namca* also show fluctuations out of the tenth rank. The collocate *namca* consistently appears in the twenty-fifth rank every year, whereas 180 does not appear in the thirtieth rank for three consecutive years from 2017 to 2019. The change in the order of the three collocates 180, *namca*, and *khi* indicates a significant decrease in the use of the new meaning since 2015.

2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
투겨	투겨	투겨	투겨	투겨	투겨	투겨	투겨	투겨	투겨
난	난	난	난	내	빅뱅	외톨이	내	내	내
키	키	키	키	키	외톨이	내	진짜	진짜	진짜
사람	내	내	내	난	시티	말	난	못	사람
전	사람	말	말	감	진짜	사람	외톨이	사람	거
워너	워너	사람	감	말	말	난	말	외톨이	말
말	전	워너	사람	사람	내	생각	못	말	못
내	말	오늘	워너	오늘	난	진짜	사람	거	새끼
생각	오늘	못	오늘	생각	베베	거	거	난	생각
남자	티켓팅	전	튀	진짜	거	키	생각	새끼	난
오늘	못	남자	못	거	사람	남자	오늘	생각	튀
년	진짜	티켓팅	진짜	못	생각	못	며	인생	인생
일	생각	튀	며	며	노래	여자	새끼	안	며
여자	튀	며	일	남자	키	며	감	오늘	오늘
튀	남자	진짜	티켓팅	튀	며	새끼	인생	며	안
며	거	생각	생각	워너	오늘	튀	튀	남자	남자
180	며	여자	거	180	취	안	남자	튀	남
거	일	일	전	여자	점쟁이	오늘	안	여자	그냥
그냥	여자	거	남자	안	튀	빅뱅	키	그냥	감
일	지금	180	본	새끼	위	180	여자	남	키
세상	180	그냥	안	티켓팅	감	인생	애	존나	애
수	그냥	지금	일	전	남자	취	수	개	여자
못	일	본	180	그냥	못	그냥	전	취	외톨이
지금	수	편	편	막	180	감	개	키	일
발언	인생	일	그냥	날	안	전	취	애	개
진짜	날	안	위	일	여자	워너	일	수	보고
정말	안	보고	콘서트	수	그냥	수	워너	보고	존나
안	보고	수	베	본	개	지금	지금	워너	계
아이콘	기본	기본	수	보고	지금	점쟁이	그냥	지금	클럽
본	샷트	인생	지금	인생	전	애	존나	점쟁이	일

Table 2 Top thirty collocates from each combined corpus

For a more detailed analysis, I examined top eleven collocates from the three corpora for each year (e.g., the 2010 combined corpus, the 2010 existing meaning subcorpus, and the 2010 new meaning subcorpus). Table 3 shows eleven collocates from the combined corpus for each year, which are categorized into three groups, with the frequency indicated below each collocate. The darkest color was assigned to overlapped collocates between the existing and new subcorpora, the second darkest color was assigned to collocates from the existing meaning subcorpus, and the third darkest color was assigned to collocates from the new meaning subcorpus. The collocates associated with the existing meaning always outnumber those associated with the new meaning. This demonstrates the dominance of the existing meaning over the new meaning throughout the ten-year period. Notably, the collocate *khi* holds the third rank for five consecutive years from 2010 to 2014 and there is only one new meaning collocate in 2015, which is ranked eleventh. In 2016, there are two collocates associated with the new meaning, which are ranked tenth and eleventh, respectively. However, from 2017 to 2019, there are no new meaning collocates among the top eleven. The changes in the order and number of the new meaning collocates indicate a decline in the use of the new meaning over time.

2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
무저 (Loser) 20792	무저 (Loser) 27824	무저 (Loser) 24781	무저 (Loser) 18129	무저 (Loser) 12519	무저 (Loser) 21228	무저 (Loser) 12487	무저 (Loser) 15452	무저 (Loser) 14718	무저 (Loser) 13872
난 (I) 2062	난 (I) 3097	난 (I) 2384	난 (I) 1317	내 (My) 863	빅뱅 (Bigbang) 2091	외톨이 (Loner/Outsider) 848	내 (My) 1017	내 (My) 1123	내 (My) 1172
키 (Stature) 1992	키 (Stature) 1812	키 (Stature) 1855	키 (Stature) 1203	키 (Stature) 827	외톨이 (Loner/Outsider) 1593	내 (My) 805	진짜 (Really) 904	진짜 (Really) 1057	진짜 (Really) 1147
사람 (Person) 1331	내 (My) 1513	내 (My) 1564	내 (My) 1196	난 (I) 819	시티 (City) 1018	말 (Word) 725	난 (I) 888	못 (Not) 964	사람 (Person) 1006
전 (I) 1305	사람 (Person) 1312	말 (Word) 1100	말 (Word) 768	감 (-) 799	진짜 (Really) 1009	사람 (Person) 637	외톨이 (Loner/Outsider) 873	사람 (Person) 911	거 (Thing) 900
위너 (Winner) 1129	위너 (Winner) 1214	사람 (Person) 1060	감 (-) 766	말 (Word) 606	말 (Word) 997	난 (I) 635	말 (Word) 831	외톨이 (Loner/Outsider) 902	말 (Word) 894
말 (Word) 1077	전 (I) 1190	위너 (Winner) 1034	사람 (Person) 765	사람 (Person) 554	내 (My) 977	생각 (Thought) 599	못 (Not) 825	말 (Word) 882	못 (Not) 775
내 (My) 897	말 (Word) 1153	오늘 (Today) 956	위너 (Winner) 733	오늘 (Today) 532	난 (I) 914	진짜 (Really) 596	사람 (Person) 780	거 (Thing) 874	새끼 (Bastard) 760
생각 (Thought) 849	오늘 (Today) 1129	못 (Not) 912	오늘 (Today) 668	생각 (Thought) 472	배배 (Bae bae) 893	거 (Thing) 574	거 (Thing) 747	난 (I) 784	생각 (Thought) 721
남자 (Man) 849	티케팅 (Ticketing) 991	전 (I) 869	뭐 (What) 611	진짜 (Really) 468	거 (Thing) 830	키 (Stature) 544	생각 (Thought) 616	새끼 (Bastard) 726	난 (I) 706
오늘 (Today) 781	못 (Not) 960	남자 (Man) 840	못 (Not) 604	거 (Thing) 446	사람 (Person) 786	남자 (Man) 519	오늘 (Today) 611	생각 (Thought) 690	뭐 (What) 573

Table 3 Eleven collocates categorized into three groups for each year

In contrast, the number of the existing meaning collocates gradually increased over the six years from 2010 to 2015. In 2016, the number of the collocates decreased from seven to five and it remained consistent from 2017 to 2018. However, in 2019, the number decreased from five to four. While the LDA analysis indicated a decrease in the use of the existing meaning in 2018, the collocation analysis shows that the decrease began in 2016. The important thing is that the year 2015 witnessed the greatest increase in the number of existing meaning collocates. This finding, along with the demoted rank of the new meaning collocate *khi* in 2015, demonstrates that the most significant change in the use of *lwuce* occurred in 2015, which agrees with the result from LDA. To sum up, the collocation analysis shows that some of the results from LDA are accurate but the following finding is not correct: in 2018, the use of the new meaning increased and the use of the existing meaning decreased. Considering this disagreement, the utilization of LDA for lexical semantic research is not currently recommended.

3.5 Discussion

The results from LDA show that there has been no semantic change in terms of the relationship between the two meanings of *lwuce*. People have more frequently used the existing meaning than the new meaning. The decrease in the use of the new meaning until 2017 demonstrates that the new meaning was actively employed when the neologism was created (in late 2009) and the related topic was a hot issue. One possible reason for the decrease might be public antipathy to the combination of the existing and new meaning. The disagreement in findings between the LDA approach and collocation analysis indicates that LDA still needs

method validation and improvement. The limitation arises from the fact that LDA does not consider word meaning. Because LDA classifies lexical items based on statistics, it is natural that its classification is different from human classification. In contrast to the LDA analysis, collocation analysis is intuitive, transparent, and clear. This is why I chose collocation analysis as a method to validate the suitability of the AI approach for linguistic research. For now, we cannot blindly rely on the AI approach and a manual analysis like collocation analysis should be required. This study suggests that collocation analysis will be able to serve as a systematic method to assess the performance of AI techniques, in particular, models trained in unsupervised learning, which lack evaluation metrics.

4. Conclusion

This study established a specific method to assess the suitability of the AI approach for semantic change research by applying a concept in language as a complex system to collocation analysis in corpus linguistics. This attempt is worth performing per se because methods associated with those two fields have not been utilized for the method validation of NLP techniques so far. This study showed that the approach of applying A-curve frequency profiles to collocation analysis works well for the method validation of the AI approach. Compared to NLP techniques with opaque internal workings, collocation analysis is clear, transparent, and evident. Therefore, employing collocation analysis can be a reliable method for the method validation of NLP techniques. The collaboration between the AI approach and collocation analysis is anticipated to encourage interdisciplinary research between the two fields, i.e., corpus linguistics and NLP.

Concerning the analysis of *lwuce*, the partial agreement of results between the LDA approach and collocation analysis demonstrates that we should not blindly believe its results. We should doubt them. The current level of NLP techniques is inferior to human analysis. An analysis relying on statistical processes is not able to explain every aspect of language because language is not numeric. Language should not be approached as if it were a mathematical problem that can be solved through the mere application of formulas. Because there are numerous variables and exceptions in language, it is essential to consider all of these factors to gain a proper understanding of language. That is where the expertise of linguists comes in. NLP engineers and computer scientists as well as analysts utilizing NLP techniques should bear this in mind. I do not argue that NLP technique-based analysis is unreliable so we should not employ it. My point is that we should use NLP techniques with greater prudence and exercise more caution in accepting their results.

References

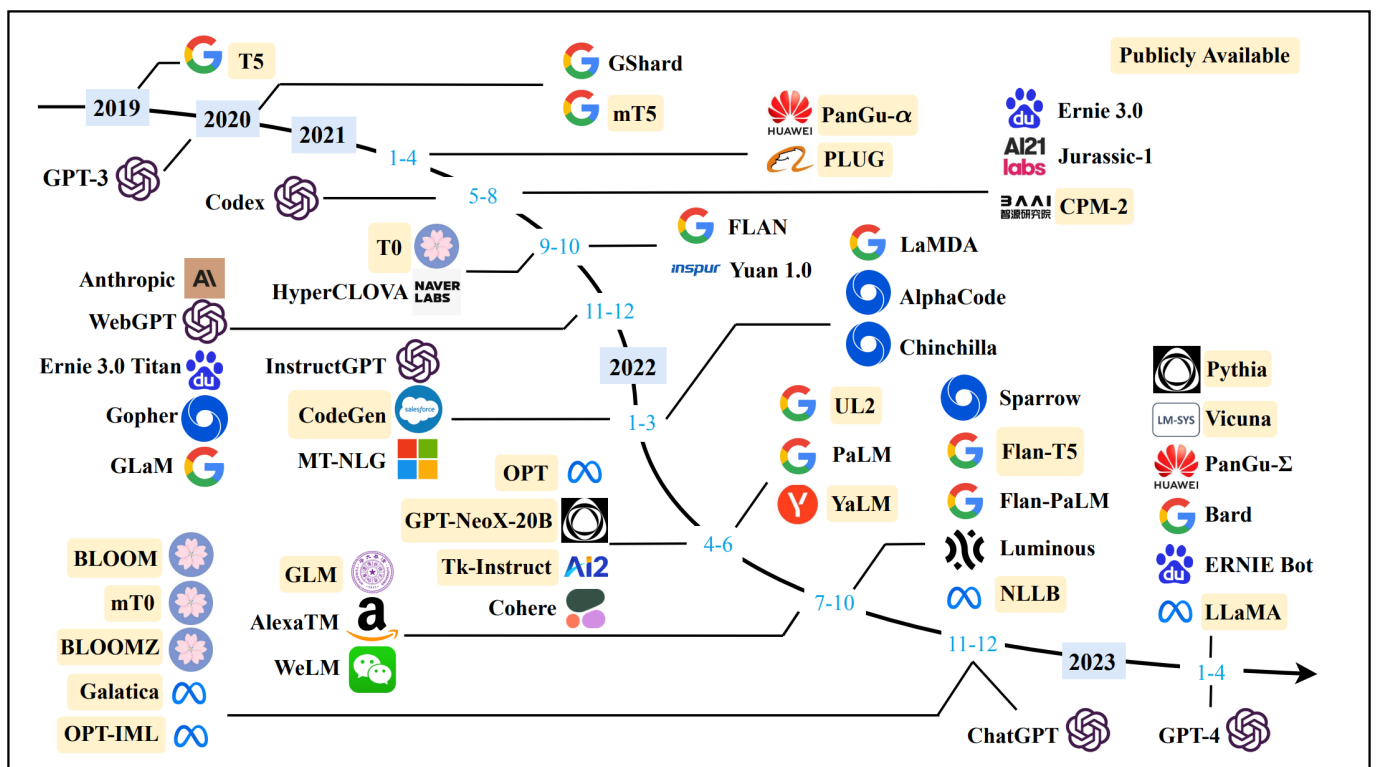
- Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3), 146-162.
- Kim, W. (2022). *Distributional corpus analysis of Korean neologisms using artificial intelligence* [Doctoral dissertation, University of Georgia].
- Kim, W. (2023). Lexical semantic analysis by means of artificial intelligence. *Proceedings of the 2023 Winter Conference of the Korean Association of Language Sciences* (pp. 349-357).
- Kretzschmar, W. A., Jr. (2009). *The linguistics of speech*. Cambridge University Press.
- Kretzschmar, W. A., Jr. (2015). *Language and complex systems*. Cambridge University Press.

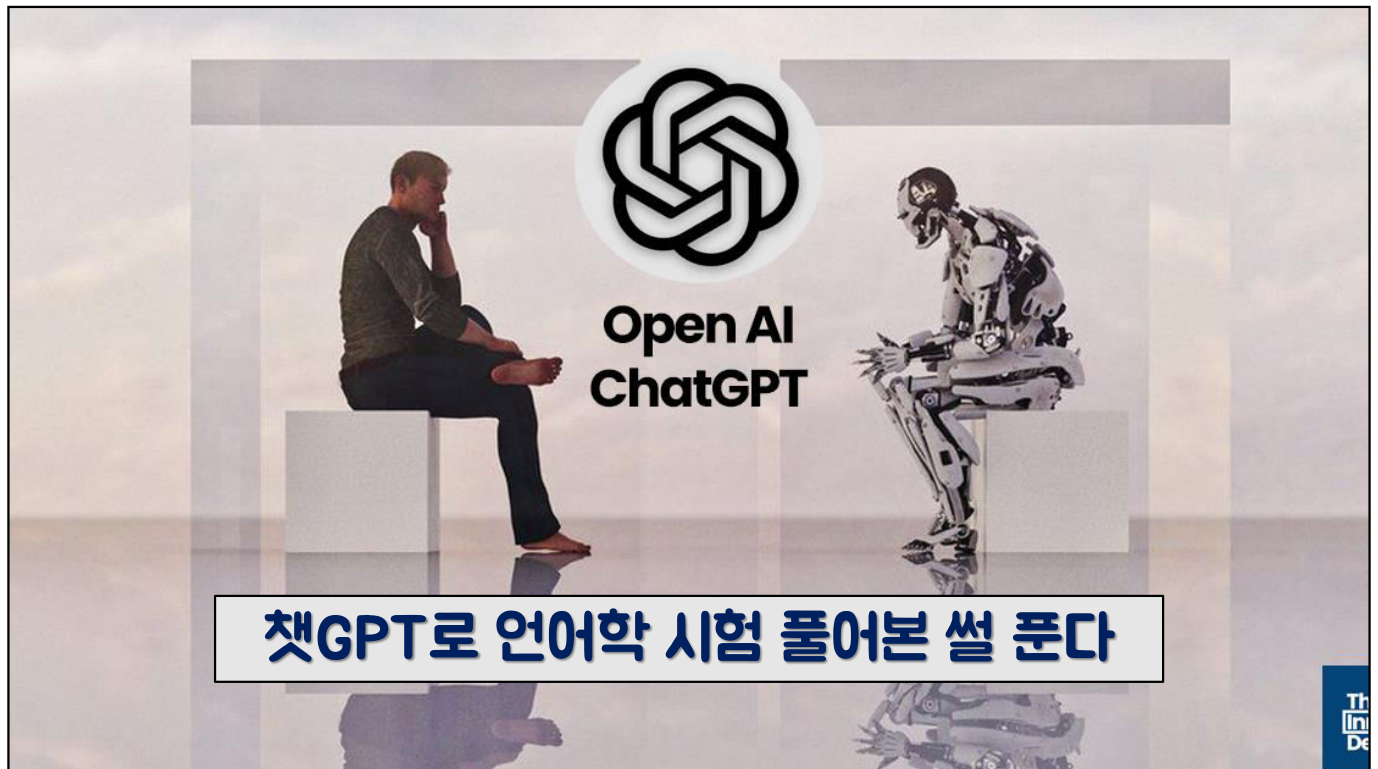
세계에서 제일 유명한
나랏말싸미 등 권에 달아
문장화로 서르스름디아니
홀씨 이런저러한 어리박
생이 나르니 저 홀배이셔도
무심내 제브들 시러피디
홀노미 하니라 내이를 위하
야 어엿비너기 새록스들여
들장를 밧이. 나기 새록스들
:히엿수빙니기. 새록스들
하리하. 나기 새록스들

챗GPT, 언어학 시험을 보다

- 협동연구 보고 -

송상현
(고려대학교 언어학과)





FAQ

A교수님: 송선생, 바빠?

me: 아닙니다. 어떤 일이신지요?

A교수님: 요 며칠 내가 챗GPT를 좀 써봤는데...

me: 아.. 네...

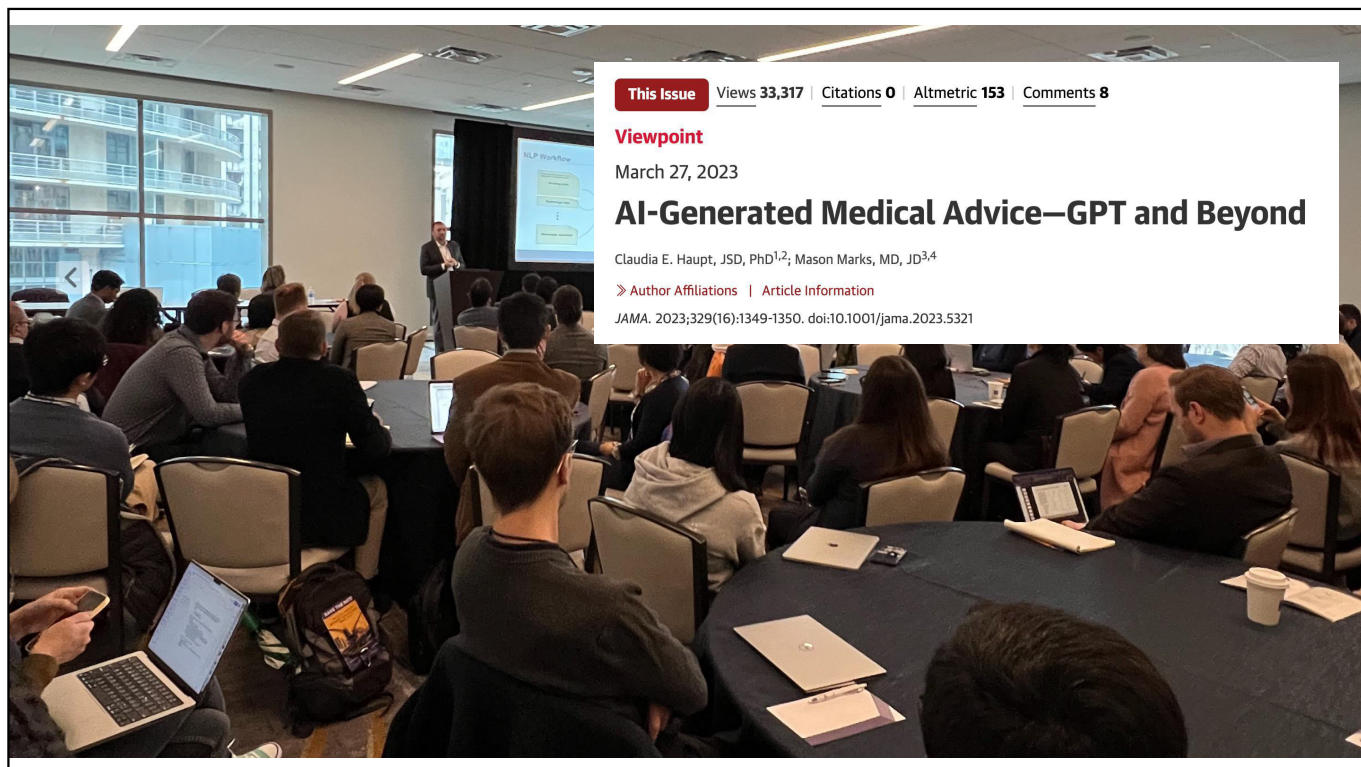
A교수님: 이제 나 수업은 어떻게 해야 할까?

주입식 교육의 종말

사용자 경험

사용자 경험(User Experience, UX)은 사용자가 어떤 시스템, 제품, 서비스를 직·간접적으로 이용하면서 느끼고 생각하게 되는 총체적 경험을 말한다.





This Issue Views 33,317 Citations 0 Altmetric 153 Comments 8

Viewpoint

March 27, 2023

AI-Generated Medical Advice—GPT and Beyond

Claudia E. Haupt, JSD, PhD^{1,2}; Mason Marks, MD, JD^{3,4}

» Author Affiliations | Article Information

JAMA. 2023;329(16):1349-1350. doi:10.1001/jama.2023.5321

DR 50세 여자가 2주 전부터 반복적으로 눈 주위와 입술이 붓는다고 병원에 왔다. 간혹 배가 아프다고 한다. 별다른 가족력은 없다. 1개월 전부터 고혈압으로 엔지오텐신전환효소억제제를 복용 중이다. 혈압 130/70 mmHg, 맥박 75회/분, 호흡 20회/분, 체온 36.5°C이다. 몸에 두드러기는 없다. 진단을 위한 가장 적절한 조치를 고르시오.

- 1) 약제를 이용한 철폐검사
- 2) 혈청 트림신분해효소 측정
- 3) 약제 중단 후 부종 소실 확인
- 4) 혈청 엔지오텐신전환효소 측정
- 5) 약제를 이용한 알레르기 피부파괴검사

GPT 이 경우, 환자의 증상을 고려하면 약물 부작용이 의심됩니다. 고혈압 약물인 엔지오텐신 전환 효소 억제제는 부종을 일으킬 수 있는 부작용이 있습니다. 따라서 가장 적절한 조치는:

3. 약제 중단 후 부종 소실 확인

DR 알레르기 가능성이 있어보이는데 5) 약제를 이용한 알레르기 피부파괴검사가 더 적절한 조치가 아닐까?

GPT 알레르기 가능성도 고려해야 하기 때문에, 환자의 증상에 대한 평가가 필요합니다. 따라서 이 경우, 5) 약제를 이용한 알레르기 피부파괴검사를 통해 알레르기 가능성을 확인하는 것이 더 적절한 조치일 수 있습니다.

DR 위 5가지 보기 중에 가장 적절한 조치는 5) 약제를 이용한 알레르기 피부파괴검사이 맞아요? 1) 약제를 이용한 철폐검사가 더 적절한지 않을까?

GPT 제가 다시 생각해보니, 환자가 알레르기과 관련된 증상을 보이고 있으므로, 1) 약제를 이용한 철폐검사가 알레르기 여부를 확인하기 위한 더 정확한 방법일 수 있습니다. 철폐검사는 약물 알레르기 반응을 확인하는데 널리 사용되는 검사입니다. 그러므로, 가장 적절한 조치는 1) 약제를 이용한 철폐검사로 생각합니다.

DR

안녕하세요. 저희 엄마가 올해 80시구요. 오랫동안 당뇨병약과 고혈압약을 복용 중입니다. 언제부터인가 조금만 걸어도 숨이차서 몇 발자국도 못걸어서 서울대병원에서 검사를 받았더니 심방세동이라는 부정맥이라고해서 11월경에 가슴에 정치를 달고 검사를 한다고 하셨습니다. 이에 대한 영양제 같은 것도 드셔도 되는지 추천해주세요.



안녕하세요. 제가 의학 전문가가 아니지만, 일반적인 정보를 제공해 드릴 수 있습니다. 심방세동(atrial fibrillation)은 심장박동이 빠르고 불규칙한 현상으로, 심방 세동 환자에게 적합한 영양제를 추천하기 전에 의사와 상의하는 것이 중요합니다. 일반적으로 심장 건강에 도움이 되는 영양제 중 일부를 소개해 드리겠습니다. 하지만 이러한 영양제를 복용하기 전에 꼭 의사와 상의하시기 바랍니다.

1. 오메가-3 지방산: 오메가-3 지방산은 심장 건강에 도움이 되며, 혈액 흐름을 개선시킬 수 있습니다. 연어, 아마씨, 호두 등의 음식에도 함유되어 있습니다.
2. 마그네슘: 마그네슘은 심장 박동을 안정화하고 혈압을 조절하는 데 도움이 됩니다. 견과류, 전분류, 채소 등 다양한 식품에서 섭취할 수 있습니다.
3. 코엔자임 Q10: 코엔자임 Q10은 심장 기능을 개선할 수 있으며, 에너지 생산과 항산화 작용을 돕습니다.
4. 비타민 D: 비타민 D는 심장 기능에 영향을 미치며, 심장 질환의 위험을 감소시킬 수 있습니다. 햇빛이나 특정 음식을 통해 섭취할 수 있습니다.

다시 한번 강조하지만, 영양제를 복용하기 전에 의사와 상의하는 것이 가장 중요합니다. 의사의 지시에 따라 적절한 복용량과 지속 기간을 결정하실 수 있습니다. 심방세동과 관련된 치료와 관리에 대해서는 전문가의 조언을 따르시기 바랍니다. 건강하세요!



나이가 들수록 심장뿐 아니라 우리 몸의 장기들은 슬슬 고장이 나기 시작합니다. 어르신들이 나이가 드시면서 이곳저곳 아프기 시작하면 자식 된 입장에서 마음이 아픈 것은 어쩔 수 없는 경우 같습니다.

현재 당뇨, 고혈압 약을 드시고 있다고 하셨는데요. 이들 약은 간이나 신장으로 대사가 되기 때문에 대사과정에서 간이나 신장에 무리를 주게 됩니다. 영양제 역시 마찬가지인데요. 아무리 좋은 영양제라 하더라도 우리 몸에서 좋은 성분만 남기고 잘 대사가 되어야 이로운 작용을 합니다. 하지만 간이나 신장이 대사 과정에서 무리를 받는다면 아무리 좋은 영양제라 하더라도 몸에 좋기는커녕 오히려 장기들만 더욱 망가지는 결과를 보일 수 있습니다.

심방세동으로 장치를 하고 검사를 진행한다고 하셨는데요. 보통 24시간 심전도를 체크하기 위해 이러한 검사를 진행하게 됩니다. 입원해서 진행할 수도 있고, 장치를 달고 집에서 생활하다 병원에 내원하여 장치를 수거한 후 판독을 할 수도 있는데요. 아마도 장치를 단다는 것은 진단을 위한 24시간 심전도 검사를 할 것으로 생각됩니다. 심방세동으로 숨이 차거나 가슴 통증이 유발되거나 하는 증상들이 자주 발생한다면 심박조율 기를 삽입할 수도 있고, 심방세동 약을 추가로 복용하게 될 수도 있습니다.

약이 추가된다면 당연히 이 역시 간이나 신장으로 대사, 배설이 되기 때문에 약이 추가될수록 몸에 무리가 되는 것도 맞습니다. 현재 상태에서는 이미 당뇨약, 고혈압 약을 드시고 있기 때문에 더 이상 무리를 주지는 않는 것이 좋겠는데요. 영양제를 추가로 복용하기보다는 몸에 좋은 음식으로 건강관리를 하시는 것이 좋을 것으로 생각합니다. 왜냐하면 심장 검사 후 추가적인 약 복용이 있을 것으로 생각되기 때문에, 간이나 신장을 좀 쉬게 하는 것이 좋겠다는 생각입니다.

하지만 이는 직접 진료 보지 않아 정확한 답변이 되지 않을 수 있으니, 반드시 어머님 다니시는 병원의 의료진과 상의하시기 바랍니다. 직접 진료 본 의료진만이 제일 정확하게 어머님 상태를 알고 있기 때문입니다. 답변이 도움이 되었기를 바랍니다. 감사합니다.

March 28, 2023

When Did GPT Graduate in Medicine?

Maurizio Balestrino, Medical Doctor | Dept. of Neurosciences (DINOEMI), University of Genoa, Italy and IRCCS Policlinico San Martino, Genoa, Italy

"Dr. Google" has already been complementing and sometimes replacing medical advice for a long time, but it has at least one clear advantage over GPT and similar AI models: its sources can be verified and independently evaluated. I can Google a search string and see with my own eyes where it took me. Obviously, it will make a big difference if it took me to the website of a respected medical association or to a blog written by a layman. The great limit (and, in my opinion, danger) of GPT is that you do not know where it derived its ...

챗GPT가 언제 PhD in Linguistics를 땀지?

Beyond the Imitation Game

- Brown, Tom, et al. "Language models are **few-shot learners**." Advances in neural information processing systems 33 (2020): 1877–1901.
- Wei, Jason, et al. "**Finetuned language models are zero-shot learners**." arXiv preprint arXiv:2109.01652 (2021).
- Kojima, Takeshi, et al. "**Large Language Models are Zero-Shot Reasoners**." arXiv preprint arXiv:2205.11916 (2022).
- Srivastava, Aarohi, et al. "**Beyond the Imitation Game**: Quantifying and extrapolating the capabilities of language models." arXiv preprint arXiv:2206.04615 (2022).

No.	Template	Accuracy
1	Let's think step by step.	78.7
2	First, (*1)	77.3
3	Let's think about this logically.	74.5
4	Let's solve this problem by splitting it into steps. (*2)	72.2
5	Let's be realistic and think step by step.	70.8
6	Let's think like a detective step by step.	70.3
7	Let's think	57.5
8	Before we dive into the answer,	55.7
9	The answer is after the proof.	45.7
-	(Zero-shot)	17.7

프롬프트에 따른 정확도 향상
(Kojima et al. 2022, 8쪽)

- **프롬프트 엔지니어**
인공지능 모델이 이해하고 반응할 수 있는 효과적인 프롬프트를 개발하고 최적화

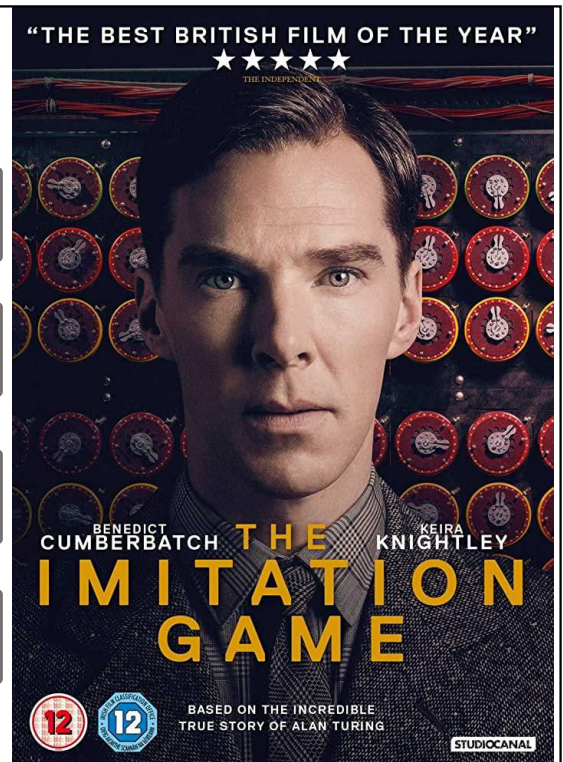
협동 연구: 튜링 테스트

통사 출제자: 김수연(세종대)
채점자: 최혜원(이화여대) / 이주은(숭실대)

음운 출제자: 오은진(이화여대)
채점자: 박선우(계명대) / 조혜선(단국대)

대조 이화여대에 재학중인 대학원생 2인

구동 송상헌 (GPT3.5 + GPT4)



용어 terminology

첫째 GPT2, GPT3, GPT3.5, GPT4
>>> 이름만 같을 뿐, 사실상 다른 모델로 간주해야 한다.

둘째 LLM: Large Language Model
>>> 얼마만큼 커야 large인가?
>>> Model이 뭔가? : history

셋째 GPT: Generative Pre-trained Transformer
>>> 생성 (생성문법, Generative Grammar) creative, production, writing, ...
>>> 사전학습
>>> Bidirectional Encoder Representations from Transformers

현황 trends

첫째 GPT4, chatGPT, RLHF(강화학습)
>>> 맛있는 냉면을 먹기 위해서는?

둘째 OpenAI >> CloseAI
>>> ChatGPT, DALL-E, Whisper
>>> 냉장고, 세탁기, TV

셋째 생성AI의 시대에 적응하지 못하면 이제 빠른 속도로 시장에서 도태된다.
>>> AI기반 기술 진보의 특징: 노빠꾸 (cf. 필름 카메라)
>>> 사람들은 이미 생성AI에 적응하기 시작했다!

단점 weakness

- 첫째** 모델 사이즈가 너무 커서, 사실상 학습 불가능
 >>> LLaMA: A100 × 2,048
- 둘째** 치명적 망각(Catastrophic Forgetting)이 발생
 >>> 다른 종류의 작업을 학습하면 이전 학습 내용을 잊어버린다.
- 셋째** 실시간성이 매우 떨어진다.
 >>> 대한민국의 대통령이 누구인가?
- 넷째** Hallucination: 뽕카 (Confabulation, 작화증)
 >>> 知之爲知之, 不知爲不知, 是知也 (논어): 모른다는 것이 무엇인지를 모른다.

질문 & 토론





과학적인 방법론으로 시장에 새로운 데이터 서비스를 제공한다는 미션으로
설립한 메트릭스는 최근 수년째 국내조사회사 **성장율 1위**를 유지하고 있습니다.
2023년에도 임직원의 힘을 모아 고객과 함께 더 큰 도전을 준비하고자 합니다.

메트릭스 그룹

- 설립년도 : 1999. 6. (2020년 물적분할)
- 매출액 : 363억 원 (2022년)
- 매출성장율 : 67% (최근 3년 성장율 60%)
- 영업이익율 : 12.7%
- 종업원수 : 140명 (2022년 말)
- 사업영역 : 시장조사, AI 데이터 구축 사업, NLU기반 언어자원 구축 사업