

소프트웨어야 놀자 아동용 처리방침 언어 수준 검증 보고서

2023. 4. 14.

서울대학교 국어교육연구소



I. 검증 방법

- 본고에서 수행한 언어 수준 검증은 “소프트웨어야 놀자 아동용 처리방침” 콘텐츠¹⁾에 대한 것으로, 크게 1) 어휘 차원과 2) 문장 차원으로 나뉘어 이루어졌으며, 이를 바탕으로 종합적인 언어 수준에 대해 검토하였다. 본고에서 사용한 구체적인 검증 기준은 다음과 같다.

1. 어휘 수준

1) 타입/토큰 비율(TTR)

- 먼저 어휘 수준에서는 텍스트에 사용된 타입(type)/토큰(token) 비율(TTR, Type/Token Ratio)을 사용하였다. ‘타입/토큰 비율’이란 텍스트에 사용된 어휘의 다양도를 뜻한다. 토큰이란 텍스트에 사용된 전체 어휘 수를 뜻하는 것으로, 중복되어 사용된 어휘를 모두 계산한다. 즉 아래와 같은 문장의 토큰 수는 7개이다. 또한 타입이란 텍스트에서 중복되지 않고 사용된 어휘 수를 뜻하는 것으로, 개별 어휘는 모두 1회씩만 계산한다. 즉 아래와 같은 문장의 타입 수는 6개이다.²⁾

타입은 중복되는 어휘를 하나의 어휘로 보는 개념입니다.

- 타입/토큰 비율은 텍스트의 이독성, 즉 텍스트를 읽고 이해하기 쉬운 정도(이성영, 2011: 170)를 설명하는 매우 중요한 요소로, 타입/토큰 비율이 높다면 해당 텍스트 내에서 중복 어휘가 적게 출현하여 어휘(토큰)의 종류가 다양하다는 것을 의미한다. 반대로 타입/토큰 비율이 낮은 경우에는 중복 어휘의 출현 빈도가 높아 어휘의 종류가 한정적이라는 의미로 해석할 수 있다.
- 결과적으로 타입/토큰 비율이 낮을수록 텍스트에서 한정적인 어휘를 중복하여 사용하고 있다는 것이므로 텍스트의 어휘 난이도가 낮다는 것을 의미하며, 텍스트의 이독성은 전반적으로 높아지는 경향을 보인다.

1) 이하에서는 ‘텍스트’로 통일하였다.

2) 한국어에서 타입/토큰을 계산할 때에는 조사와 어미는 제외한다.

2) 어휘 등급

- 다음으로 텍스트에 사용된 어휘의 등급을 분석하였다. 어휘 등급은 국어교육 연구에서 가장 많이 사용되고 있는 김광해(2003)의 어휘 등급을 활용하였다. 김광해(2003)는 그간 선행 연구에서 제시된 어휘 자료를 메타(meta) 계량³⁾이라는 방법을 활용하여 선정 및 평정한⁴⁾ 연구로 총 238,010개의 어휘를 선정하였으며, 이를 1등급(기초어휘)에서 7등급(누락어)으로 평정한 바 있다. 그간 이루어진 몇몇의 어휘 선정 연구가 같은 연구 목적을 가지고 진행되었는데도 크게 상이한 경우가 많아 객관성을 담보하기 어려웠던 반면, 김광해(2003)는 메타 계량의 방법을 활용함으로써 객관적이고도 포괄적인 교육용 어휘를 선정 및 평정하였다는 데 그 의의가 있다. 김광해(2003)에서 제시한 어휘 등급은 다음 <표 1>과 같다.

어휘량	누계	국어교육용		비고
		등급	개념	
1,845	1,845	1	기초 어휘	소사전 중사전
4,245	6,090	2	정규 교육 이전	
8,358	14,448	3	정규 교육 개시~사춘기 이전, 사고도구어 일부 포함	
19,377	33,825	4	사춘기 이후~급격한 지적 성장, 사고도구어 포함	
32,946	66,771	5	전문화된 지적 성장 단계, 다량의 전문어 포함	
45,569	112,340	6	저빈도어: 대학 이상. 전문어(기존 계량 자료 등장 어휘 + 누락어 14,424어 추가)	
125,670	238,010	7	누락어: 분야별 전문어. 기존 계량 자료 누락어휘	

<표 1> 김광해(2003)의 국어교육용 어휘 등급(김광해, 2003: 27)

- 이 중 1~2등급 어휘는 초등학교 교육을 받기 전 모어 화자가 습득하게 되는(또는 습득해야 되는) 어휘이며, 3등급 어휘는 초등학교 학습자가 습득하게 되는 어휘를 뜻한다. 다음으로 4등급 어휘는 중학교 학습자가 습득하게 되는 어휘와, 5등급 어휘는 고등학

3) “이 방법은 더 높은 객관성을 확보하기 위하여 구상된 것으로서, 직접 말뭉치를 계량하는 것이 아니라 기존에 계량 처리된 자료들을 충분히 구하여 그 분포 상황과 자료의 타당도를 함께 고려하면서 비교함으로써 중요도를 정하는 방법을 말한다(김광해, 2003: 22).”

4) 어휘 선정은 국어교육용 어휘 목록을 구축하는 것을 말하며, 어휘 평정은 선정된 어휘 목록을 중요도에 따라 등급화하는 것을 말한다(김광해, 2003: 11).

교 학습자가 습득하게 되는 어휘와 대략적으로 일치한다.

- 본고에서 대상으로 하는 텍스트는 전체 이용자를 대상으로 제공하는 것으로서 만 14세 미만의 어린이가 이해하기에도 무리가 없는 텍스트여야 하기에, 중학교 학습자를 포함하여 초등학교 학습자까지 이해하기에 무리가 없는 언어 수준으로 작성될 필요가 있다.

2. 문장 수준

- 문장당 어절 수는 문장의 길이를 나타내는 지표로, 문장의 길이 또한 이독성에 영향을 미치는 매우 중요한 요소이다. 문장당 어절 수에 따른 텍스트의 수준을 수치적으로 말하기는 어려우나, 문장당 어절 수가 많을수록 텍스트가 이해하기 어려워지며, 즉 이독성이 떨어진다.
- 따라서 본 텍스트의 경우 만 14세 미만의 어린이도 이해하기에 어려움이 없도록 가급적 문장을 나누어 짧은 문장으로 글을 작성할 필요가 있으며, 의미 전달에 반드시 필요하지 않은 부속 성분(관형어와 부사어 등)은 삭제하여 간결한 문장 표현을 할 필요가 있다.

II. 검증 결과

1. 어휘 수준

1) 타입/토큰 비율

- 텍스트에 제시된 전체 어휘 토큰은 681개로,⁵⁾ 본 텍스트에는 총 681개의 어휘가 사용되었다. 어휘 타입은 238개로, 텍스트의 타입/토큰 비율은 34.9%⁶⁾이다. 즉, 전체 텍스트에서 약 34.9%의 단어가 처음 등장하는 어휘이며, 나머지 65.1%의 단어는 앞서 등장한 어휘이다.

5) 네이버, 웨일 등 고유 명사나 숫자, 알파벳 등은 제외하였다. 또한, 은전한닢 프로젝트에서 개발한 Mecab-ko 형태소 분석기를 사용하여 분석을 진행하되, 일부 분석 오류에 대해서는 다시 수정하는 과정을 거쳤다.

6) 모든 숫자는 소수점 둘째 자리에서 반올림하였다.

- 이와 관련하여 교과서의 타입/토큰 비율에 대한 선행 연구를 참조할 수 있다. 고등학교 1학년 영어 교과서의 타입/토큰 비율을 분석한 이용훈 외(2015)에 따르면, 타입/토큰 비율은 출판사별로 20~28% 사이로 나타났다.
- 아울러 2007 개정 교육과정 및 2009 개정 교육과정 초등학교 교과서의 타입/토큰 비율을 분석한 이기연(2015)에 따르면 초등학교 교과서의 타입/토큰 비율은 교과목에 따라 차이는 있으나 2007 개정 교육과정의 경우 평균 22% 정도로 나타났으며, 특히, 수학 교과를 제외하면 약 23.7%의 타입/토큰 비율을 가지는 것으로 확인되었다. 또한, 2009 개정 교육과정의 경우 5~6학년군의 교과서가 포함되지 않았는데도 불구하고 30%의 타입/토큰 비율을 보였다.
- 따라서 해당 텍스트의 타입/토큰 비율은 34.9%로 이용훈 외(2015), 이기연(2015) 등 선행 연구에 비해 다소 높게 나타나 텍스트에 사용된 어휘가 상대적으로 다양하다고 볼 수 있다. 즉, 어휘 다양도의 측면에서는 해당 텍스트의 난도가 다소 높을 수 있다.
- 다만, 타입/토큰 비율이 텍스트의 크기가 작을수록 커진다는 점을 고려한다면, 그리고 교과서의 어절 수에 비해 해당 텍스트의 어절 수가 매우 적다는 점을 고려한다면 해당 텍스트의 이독성은 양호한 것으로 판단할 수 있다.

2) 어휘 등급

- 텍스트에 제시된 전체 어휘(토큰 기준)의 등급별 분포는 아래와 같다.

어휘 등급	단어 수(%)	비고
1등급	313(46.0%)	개인, 위하다, 있다, 경우 등
2등급	155(22.8%)	정보, 이용, 놀다, 서비스 등
3등급	114(16.7%)	소프트웨어, 수집, 회원, 관리 등
4등급	43(6.3%)	동의, 제공, 가입, 방침 등
5등급	46(6.8%)	이용자, 삭제, 대리인, 로그인 등
6등급	0(0.0%)	-
등급 없음	10(1.5%)	아이디, 버전, 닉네임, 오프라인 등
총계	681	
평균 어휘 등급	2.04	* 등급 없음 단어 제외

<표 2> 텍스트의 토큰 기준 어휘 등급

- 분석 결과 1등급 어휘가 약 46.0%로 가장 많은 분포를 보이고 있으며, 다음으로 2등

급 어휘는 약 22.8%이다. 3등급 어휘는 약 16.7%로, 전체 어휘의 약 85.5%가 초등학교생에게 적합한 1~3등급 어휘로 구성되어 있다. 다음으로 4등급 어휘가 6.3%로 나타나, 전체 어휘의 91.8% 정도가 중학생에게 적합한 1~4등급 어휘로 구성된 것을 확인할 수 있었다. 또한, 고등학생에게 적합한 5등급 어휘는 6.8%로 나타났다. 전체 평균 어휘 등급은 2.04등급으로 확인되었다.

- 이와 관련하여 초등학교 3학년 2학기 교과서의 어휘 등급을 비교한 이성영(2011)에서는 다음과 같은 결과를 제시하고 있다. 초등학교 3학년은 10세(만 8~9세)로 10세 아동 대상의 교과서의 어휘 수준이 대체로 1.5~1.7등급 정도임을 알 수 있다.

교과서	어절 수	등급의 합	평균 등급
국어	644	986	1.53
사회	634	1094	1.73
과학	650	1054	1.62

<표 3> 교과서별 전체 어휘 비교 결과(이성영, 2011: 183)

- 따라서 해당 텍스트의 어휘 수준은 이성영(2011)의 결과보다는 다소 높게 나타나, 초등학교 3학년 교과서에 비해서는 상대적으로 높은 난도의 어휘가 일부 사용된 것으로 확인되었다. 다만, 90% 이상의 어휘들이 중학생에게 적합한 1~4등급 어휘로 구성되어 있고, 특히, 초등학생에게 적합한 1~3등급 어휘가 대부분을 차지한다는 점을 고려했을 때, 만 14세 미만의 어린이가 이해하기에도 큰 무리가 없을 것으로 판단된다.

2. 문장 수준

- 해당 텍스트의 문장 수는 제목 등 비문장 형식을 제외하고, 총 47개로 나타났다. 총 어절 수는 571개이다. 따라서 문장당 어절 수는 12.1(571÷47)개로 나타났다.
- 한편, 이성영(2011)에 따르면 초등학교 3학년 2학기 교과서의 문장당 평균 어절 수는 다음과 같이 나타났다. 만 8~9세 대상 교과서의 문장 길이는 약 9~11어절 수준임을 알 수 있다.

교과서	어절 수	문장 수	문장당 어절 수
국어	644	73.2	8.80
사회	634	64.5	9.83
과학	650	59.5	10.92

<표 4> 교과서별 전체 어휘 비교 결과(이성영, 2011: 184)

- 또한, 초등학교 및 중학교 국어 교과서의 정보 텍스트를 양적으로 분석한 길호현 외 (2022)에서는 초등학교 2학년 교과서의 문장당 평균 어절 수가 8.75어절, 3학년의 경우 9.45어절, 4학년의 경우 10.16어절, 5학년의 경우 12.26어절, 6학년의 경우 13.65어절을 밝혔다. 중학교 1학년 교과서의 경우에는 12.33어절, 중학교 2학년의 경우 13.33어절의 문장당 평균 어절 수를 가지는 것으로 확인되었다.
- 결론적으로 문장당 어절 수에 따른 이독성을 정확한 수치로 말하기는 어려우나, 해당 텍스트는 하나의 문장당 평균적으로 12.1개의 어절이 사용되어, 문장의 길이가 만 14세 미만의 어린이의 수준에 적합한 것으로 나타났다. 특히, 해당 텍스트가 법률적인 내용을 다루고 있는 텍스트라는 점, 그리고 일반적으로 법률 텍스트의 문장 길이가 길다는 점을 고려한다면 해당 텍스트의 문장당 어절 수는 양호한 것으로 판단된다.

3. 전반적인 언어 수준

- 본 아동용 콘텐츠의 전반적인 언어 수준은 다음과 같이 요약될 수 있다. 우선 어휘 수준의 측면에서 타입/토큰의 비율이 34.9%로 다소 높기는 하지만, 어휘 등급의 수준에서 약 85.5%의 단어가 초등학생 아동에게 적합한 1~3등급 어휘, 약 91.8%의 단어가 중학생 아동에게 적합한 1~4등급 어휘로 구성되어 있다. 또한 전체 평균 어휘 등급은 2.04등급으로 나타나, 어휘 수준에서 본 콘텐츠는 만 14세 미만의 어린이에게도 적합하다고 볼 수 있다.
- 다음으로 문장 수준에서 문장 당 어절 수가 12.1어절로 나타나 문장의 길이가 양호한 것으로 나타났다. 특히, 본 콘텐츠가 법률적인 내용을 다루고 있는 텍스트라는 것을 고려한다면, 해당 콘텐츠의 문장당 어절 수는 만 14세 미만의 어린이에게도 적합한 수준이라고 판단할 수 있다.

참고 문헌

- 길호현·류수경·박태준·최소영(2022), 교육용 텍스트의 이독성 평가를 위한 양적 준거 연구 : 정보 텍스트를 중심으로, 독서연구 63, 한국독서학회, pp.9-43.
- 김광해(2003), 등급별 국어교육용 어휘, 박이정.
- 이기연(2015), 초등학교 교과서 수록 어휘 양상 연구, 학습자중심교과교육연구 15(6), 학습자중심교과교육학회, pp.669-689.
- 이성영(2011), 읽기(독서)에서의 교육 내용 위계화: 초등 교과서의 이독성 비교 연구, 국어교육학연구 41, 국어교육학회, pp.169-193.
- 이용훈·이규형·김하응(2015), 표준화된 타입-토큰 비율, 어휘성장곡선, 그리고 영어교재분석, 영어학연구 21(2), 한국영어학학회, pp.1-20.