Applied Language Learning

Research

Beyond News and Documentaries: Developing a Corpus-Based Lexical Resource for Informal North Korean Speech

Mi Hye Lee
Defense Language Institute Foreign Language Center, Monterey, California

Myoyoung Kim
Defense Language Institute Foreign Language Center, Monterey, California

By conceptualizing North Korean (NK) news and NK comedy talk shows as representing distinct Target Language Use domains (Bachman & Palmer, 2010), this study investigates lexical variation across formal and informal registers in contemporary NK discourse and considers whether differences in communicative purpose and discourse conditions are reflected in measurable lexical patterns. Two groups of video clips were compiled, manually transcribed, and normalized: an informal spoken corpus based on 90 minutes of seven NK comedy talk show episodes and a formal spoken corpus based on 90 minutes of NK news broadcasts. The two corpora datasets were analyzed with a focus on part-of-speech (POS) distributions and lexical frequency. The results reveal clear register-based differences in lexical distributions as well as POS categories. In particular, nouns dominate the news corpus, whereas the comedy corpus shows a higher proportion of adverbs, indicating more descriptive and emotionally expressive language use. The comparison was intended not merely to document register contrasts, but to evaluate whether reliance on a single text type, particularly formal discourse, may lead to a skewed lexical representation in instructional materials. In doing so, the analysis provides empirical evidence relevant to questions of domain representativeness in NK language education and highlights the potential limitations of narrowly defined instructional input.

Keywords: *North Korean, Corpus Analysis, Formal and Informal Register, Token and Type, Parts of Speech*

INTRODUCTION AND BACKGROUND

Although North Korean (NK) has often been perceived as largely similar to South Korean (SK), more than 80 years of division between the two have led to significant divergence across multiple linguistic areas, including morphology, syntax, and phonology. Most research on the North Korean language, conducted primarily by South Korean scholars, expanded significantly after North Korea's official announcement of its *Revised Korean Language Education Guidelines* in 2013 (Eum & Seo, 2021; Kim, J., 2015; Kim, N., 2024; Oh, 2015). Prior to 2013, research tended to focus on identifying and comparing linguistic differences between SK and NK. In contrast, after the release of North Korea's national language curriculum in 2013, scholarly attention of SK researchers shifted toward exploring ways to restore linguistic homogeneity between the two languages in preparation for reunification.

Among these efforts, vocabulary-related research has been particularly active, reflecting the growing recognition that vocabulary will play a crucial role in communication in the future. A survey conducted by the National Institute of the Korean Language on the settlement experiences of NK defectors in South Korea further supports the existence of a linguistic gap. Defector respondents identified language, particularly SK vocabulary and pronunciation, as the most challenging aspect of adapting to SK society (Oh, 2015). Testimonies indicating that communication was difficult due to unfamiliar vocabulary clearly demonstrate both the importance of vocabulary learning and the substantial linguistic gap between SK and NK.

At the Defense Language Institute Foreign Language Center (DLIFLC), Korean School students are required to develop proficiency in both SK and NK for their future professional assignments. Accordingly, instructors have consistently sought effective strategies to enhance learners' NK proficiency. Students typically build a foundation in SK vocabulary and grammatical structures before transitioning to NK. In this context, focused vocabulary instruction has been considered one of the most effective means of supporting NK acquisition, as the primary challenge often lies not in syntactic differences but in lexical variation in word choice across contexts, speakers, regions, social groups, and communicative situations.

However, the impact of vocabulary instruction depends critically on the representativeness of the lexical items selected. This concern is particularly relevant in the NK materials used in DLIFLC. Current NK instructional materials rely predominantly on highly formal, government-produced texts such as news broadcasts and official publications. These texts represent a narrowly defined institutional topical domain characterized by scripted discourse, ideological positioning, and conventionalized lexical and grammatical patterns. Even spoken news broadcasts are pre-scripted and informational rather than interactive, reinforcing their classification as formal register. By contrast, informal spoken discourse—featuring spontaneous interaction, colloquial expressions, and pragmatic negotiation—is minimally represented in existing DLIFLC materials. This raises concerns of domain underrepresentation (Bachman & Palmer, 2010). Learners exposed primarily to formal registers may lack exposure to lexical and structural features typical of everyday communication.

Achieving language proficiency requires sustained exposure to the full range of language use situations that learners are likely to encounter in real life. All languages operate across a continuum, extending from highly regulated formal forms to informal forms embedded in everyday life. Although formal language is often associated with written discourse and informal language with spoken communication, this binary distinction is overly simplistic. Register differences cannot be reduced to modality alone. Rather, they emerge from communicative purpose, participant roles, social relationships, situational constraints, and sociocultural norms governing language use. A comprehensive approach to foreign language (FL) education must therefore recognize that proficiency entails not only grammatical accuracy, but also the ability to navigate shifting communicative contexts with appropriate lexical, structural, and pragmatic resources.

This perspective is captured in the concept of the Target Language Use (TLU) domain, proposed by Bachman and Palmer (2010). A TLU domain refers to the set of real-world communicative situations in which language is used and for which language ability is evaluated. It encompasses communicative purposes, language users, settings, tasks, discourse conditions, and the linguistic and pragmatic demands these contexts impose. Within assessment theory, the TLU domain serves as a reference point for ensuring domain representativeness and authenticity. Hence, instructional materials should reflect the communicative realities learners are expected to manage outside the classroom. Although originally developed for language testing, the TLU framework offers a powerful lens for examining instructional materials more broadly. If materials systematically represent only a limited portion of relevant communicative domains, learners may develop competence in that restricted domain while remaining underprepared for others.

Despite the pedagogical importance of such distinctions, research on NK vocabulary remains limited in scope and methodology. Previous studies have examined lexical differences between the two Koreas (Cheong, 2021; Kim, J., 2015; Lee, 2007), proposed the development of NK vocabulary textbooks grounded in intercultural perspectives (Eum & Seo, 2021), explored approaches to NK vocabulary education treating NK as a regional variety (Kim, N., 2024), and suggested the development of instructional materials emphasizing comprehension of NK (Oh, 2015). To date, no study has specifically examined differences between formal and informal registers within NK, which constitutes the primary focus of the present study. This may be because, for native speakers of SK, differences between formal and informal NK registers seldom create substantial comprehension barriers, particularly in listening. By contrast, for foreign learners of NK, such as the students we have at DLIFLC, even minor phonological or grammatical variations may be perceived as distinct lexical and structural features requiring deliberate acquisition.

From a broader research context, corpus-based register studies in Korean and other languages, particularly those adopting multidimensional approaches (e.g., Biber, 2001, 2004; Kim & Biber, 1994; Reppen, 2001), have emphasized grammatical morphology, information density, and global text-structural features, including lexical bundle classification. These studies demonstrate that formal discourse tends toward nominalization and high informational load, whereas informal discourse typically exhibits greater use of modifiers and interactional markers. However,

less attention has been given to lexical distributional patterns at the level of part-of-speech (POS) frequency within a single language variety. To date, no study has systematically examined corpus-based lexical differences between formal and informal registers in NK.

While corpus-based research has primarily advanced our understanding of linguistic patterns, its implications extend beyond theoretical inquiry into language pedagogy. Particularly, corpus-based language data plays an increasingly central role in the development of instructional materials (Curry & Mark, 2024; Moser, 2020), especially in contexts where authenticity and representativeness are essential (Latham, 2025; Li et al., 2025). This empirical foundation supports more principled decisions in vocabulary selection, ensuring that instructional content reflects actual language use within defined communicative domains. Moreover, quantitative comparisons between formal and informal discourse allow instructional material developers to detect imbalances and disparities in lexical and structural representation. In this way, corpus analysis serves not only as a descriptive tool but also as a means of evaluating whether instructional materials adequately represent relevant TLU domains, thereby promoting closer alignment between classroom instruction and real-world language use.

In the present study, NK news broadcasts are classified as formal because of their scripted, institutionalized, and information-dense nature, whereas NK comedy talk shows are classified as informal because they involve spontaneous, interactive exchanges employing everyday colloquial language. Conceptualizing these genres as distinct TLU domains allows register differences to be examined within a principled framework rather than as purely stylistic contrasts.

Building on the theoretical and pedagogical framework outlined above, the following section presents a systematic comparison of formal and informal NK texts through a corpus-based analysis of their lexical features.

Research Questions

The study is guided by two primary research questions:

RQ #1. Are the lexical components used in NK formal and informal texts similar or different in terms of POS distribution and frequency of lexical items?

This question addresses the extent to which register variation manifests at the structural level of lexical organization. If formal discourse is characterized by higher nominal density and informational load, while informal discourse exhibits greater use of modifiers, interactional markers, or other lexical categories, such differences should be observable in corpus-based distributions. By examining POS patterns and lexical frequency profiles across the two corpora, the study aims to provide systematic evidence of intra-varietal register variation within NK.

RQ #2. If differences exist, in what ways are these differences reflected, or not reflected, in current NK language instructional materials?

This question directly connects corpus findings to pedagogical practice within the TLU framework. If instructional materials are predominantly drawn from formal news discourse, they may disproportionately represent lexical features associated with that domain while underrepresenting those typical of informal interaction. Evaluating this alignment allows for a more principled discussion of whether NK instruction adequately prepares learners for a broader range of communicative contexts.

Together, these research questions integrate corpus analysis with TLU-based considerations of domain coverage, positioning the study at the intersection of register research and instructional material development.

METHOD

Data Sources

Approximately 90 minutes of NK news clips and an equal amount of comedy talk show data (seven videos in total) were transcribed and used to build two comparable corpora (see Appendix A for detailed video information). For the analysis of informal speech, comedy talk shows (called 재담, 만담, or 촌극 in North Korea) were selected because they focus on everyday life topics and avoid overt political content, thus making them suitable for analyzing informal language use. While NK news broadcasts are relatively accessible through official state-run broadcasting channels (e.g. Korean Central Television: KCTV; and North Korean News and Media Websites), comedy talk show materials are extremely limited. They were initially obtained from a North Korean defector and are now confirmed to be publicly available on YouTube. The links are provided in Appendix A.

Building Corpora

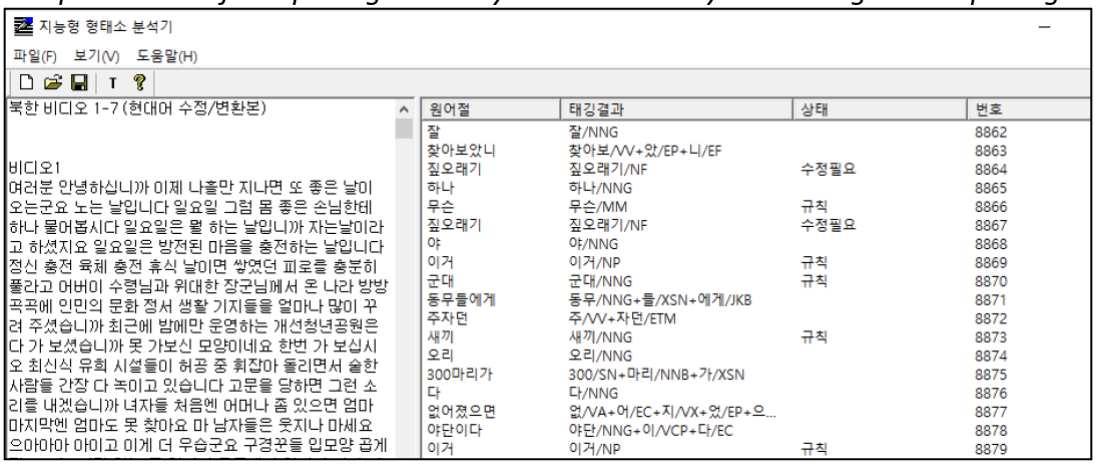
Since both sets of data consisted of spoken materials, a transcription process was required for corpus building. Although existing transcription tools perform reasonably well for standard SK speech, their performance deteriorates markedly when applied to NK speech. The error rate was particularly high in NK comedy talk shows, where diverse intonation patterns, informal expressions, and nonstandard lexical items are common. This reduced accuracy appears to stem from the limited availability of NK speech data for model training. Given these substantial transcription errors, automatic speech recognition was not deemed sufficiently reliable for the purposes of this study. Instead, all recordings were transcribed manually, reviewed in consultation with a North Korean defector, and cross-checked to ensure accuracy and consistency across the dataset. After transcription, colloquial contractions and spoken expressions were normalized into standard South Korean forms, and spacing and punctuation were carefully reviewed and adjusted before finalizing the raw corpus. For this example, refer to Appendix B.

Morphological Analysis and Correction Process

The two raw corpora, news and comedy talk shows, were first automatically annotated using the Intelligent Morphological Analyzer provided by the National Institute of the Korean Language (NIKL) through its Language Information Sharing Service (https://kcorpus.korean.go.kr/). The initial automatic annotation results are as follows.

Figure 1

Sample Results of Morphological Analysis Generated by the Intelligent Morphological Analyzer



After the initial automatic annotation, manual correction as post-editing was conducted to refine the results and compile the word lists. Figure 2 presents the automatic parsing output for the same transcript shown in Figures 1. Instances of incorrect morphological analysis are highlighted in yellow below. These errors were removed from the dataset prior to analysis.

Figure 2

Example of Automatic Annotation Output and Manual Correction

	A	B	C	D	E	F
1	순번	어절	어절 형태소 분석	비고		
2	1	비디오1	비디오/NNG+1/SN	NULL		
3	2	여러분	여러분/NP	NULL		
4	3	안녕하십니까	안녕/NNG+하/XSA+시/EP+ㅂ니까/EC	NULL		
5	4	이제	이제/MAG	규칙		
6	5	나흘만	나흘/NNG+만/JX	NULL		
7	6	지나면	지나/VV+면/EC	규칙		
8	7	또	또/MAJ	규칙		
9	8	좋은	좋/VA+은/ETM	규칙		
10	9	날이	날/NNG+이/JKS	NULL		
11	10	오는군요	오/VV+는군요/EC	NULL		
12	11	노는	놀/VV+는/ETM	규칙		
13	12	날입니다	날/NNG+이/VCP+ㅂ니다/EC	NULL		
14	13	일요일	일요일/NNG	규칙		
15	14	그럼	그럼/MAG	규칙		
16	15	몸	몸/NNG	규칙		
17	16	좋은	좋/VA+은/ETM	규칙		
18	17	손님한테	손님/NNG+한테/JKB	NULL		
19	18	하나	하나/NR	NULL		
20	19	물어봅시다	묻/VV+어/EC+보/VX+ㅂ시다/EC	NULL	물어보다 과분석	
21	20	일요일은	일요일/NNG+은/JX	NULL		
22	21	뭘	무엇/NP+을/JKO	규칙		
23	22	하는	하/VV+는/ETM	NULL		
24	23	날입니까	날/NNG+이/VCP+ㅂ니까/EC	NULL		
25	24	자는날이라고	자는날이라/NF+이/VCP+고/EC	수정필요	띄어쓰기 분석 오류	
26	25	하셨지요	하/VX+시/EP+었/EP+지요/EF	NULL		
27	26	일요일은	일요일/NNG+은/JX	NULL		

RESULTS AND DISCUSSION

Through the above procedures, the total number of words (token, total word occurrences including duplicates), unique word types (unique lexical items excluding repetitions) were extracted. The results are summarized below followed by a discussion.

Table 1
Token and Type Counts of the NK News and Comedy Talk Show Data

	Token	Type
NK News (Formal Style)	14,415	2,549
NK Comedy Talk Shows (Informal Style)	15,567	2,437
Total	29,982	4,986

In addition to token and type counts, the counts for each POS and function words were also extracted. A total of 31 POS categories were identified in the news corpus and 33 in the comedy talk show corpus (see Appendix C for the complete results of both genres). Although the two corpora show minimal differences in overall type counts and the number of POS categories, the following section examines how the distribution of POS categories differs between the two corpora.

Nouns and Adnominal Items

Table 2 presents the type and token counts for nouns and particles attached to nouns. The percentages are calculated relative to the total number of types and tokens reported in Table 1 above.

Table 2
Token and Type Counts of Nouns and Adnominal Items in NK News and Comedy Talk Shows

Part of Speech	Token (Frequency)		Type	
	News	Comedy Talk Show	News	Comedy Talk Show
Common noun	3,821 (26.51%)	2,790 (17.92%)	1,308 (51.31%)	937 (38.45%)
Proper noun	247 (1.71%)	115 (0.74%)	98 (3.84%)	71 (2.91%)
Adnominal particle	499 (3.46%)	128 (0.82%)	1 (0.04%)	1 (0.04%)
Adnominal ending	1,026 (7.12%)	600 (3.85%)	10 (0.39%)	17 (0.70%)

Common nouns occur in a significantly higher proportion in NK news than in comedy talk shows, in both token and type counts. Among the total 2,549 word types, more than half (1,308, or 51.31%) were common nouns. In terms of frequency, over one-fourth of all tokens (26.51%) were also common nouns. This indicates that a wide variety of nouns are used in news broadcasts,

reflecting the genre's tendency to repeatedly present information about events and to emphasize government policies, particularly in North Korea's case.

For *proper nouns*, while the number of types showed little difference between news and comedy talk shows, their frequency in news was more than twice as high (1.71% vs. 0.74%). This is because news reports frequently include names of people, places, and institutions.

The *adnominal case particle* -의 (*possessive*), which appeared only once as a type in both corpora, showed a frequency of about four times higher in news (3.46%) than in comedy talk shows (0.82%). This can be attributed to the frequent use of noun phrases expanded using -의 (*possessive*) in news texts, whereas in comedy talk shows, short and colloquial expressions often assumed or omitted -의 (*possessive*). Korean examples in this paper are presented in the form of 'a Korean lexical word' followed by its English translation in parentheses. The meanings of grammatical morphemes are italicized to distinguish them from lexical categories. For example, 많이 (*a lot*) vs. -의 (*possessive*).

Similarly, *adnominal endings* appeared more than twice as frequently in news as in comedy talk shows (7.12% > 3.85%), despite similar type counts. This reflects the characteristics of news discourse, which tends to compress information through modifying clauses, whereas comedy talk shows prefer direct description or sequential narration rather than syntactic compression.

In summary, NK news demonstrates a dominant use of common nouns and adnominal particles or modifiers, highlighting its noun-centered and information-focused sentence structures designed for clear and forceful information delivery. This finding aligns with the idea that news discourse is characterized by a predominance of noun-based lexical elements, as reported by Metang and Narathakoon (2025) in their corpus-based study of online news, where they found that nouns, along with noun phrases, serve key informational functions in news texts.

Verbs, Adjectives, and Adverbs

Table 3

Token and Type Counts of Verbs, Adjectives, and Adverbs in NK News and Comedy Talk Shows

Part of Speech	Token (Frequency)		Type (Frequency)	
	News	Comedy Talk Shows	News	Comedy Talk Shows
Verb	1,586 (11.00%)	1,940 (12.46%)	487 (19.11%)	428 (17.56%)
Adjective	520 (3.61%)	575 (3.69%)	165 (6.47%)	136 (5.58%)
General adverb	440 (3.05%)	907 (5.83%)	128 (5.02%)	211 (8.66%)

Verbs

As seen in Table 3, in the case of verbs, the *type* count was higher in the news corpus (487 > 428), but the *token* frequency showed the opposite pattern, comedy talk shows contained more verb tokens (1,940 > 1,586). This means that although news employs a wider range of verb types, comedy talk shows repeat the same verbs more frequently. To examine which verbs were most frequently repeated, the top 30 verbs by frequency were extracted from each genre, revealing several noteworthy distinctions.

The top 30 most frequent verbs in the News and Comedy Talk Show corpora are listed below. Verbs are ordered by frequency within each corpus, and those highlighted in red indicate overlap between the two corpora.

News: 하다 (to do), 경애하다 (to respect), 위하다 (to favor), 심다 (to plant), 보다 (to watch), 만들다 (to make), 가다 (to go), 높이다 (to raise), 찾다 (to find), 받다 (to receive), 품다 (to embrace), 진행하다 (to proceed), 나서다 (to come forth), 생산하다 (to produce), 진행되다 (to be proceeded), 가지다 (to have), 구리다 (to be filthy), 맞다 (to be correct), 수행하다 (to perform), 보내다 (to send), 좋아하다 (to like), 가르치다 (to teach), 들다 (to lift), 일으키다 (to cause), 올리다 (to raise), 나다 (to come out), 이바지하다 (to contribute), 모르다 (to not know), 돌아보다 (to recollect), 간직하다 (to keep)

Comedy Talk Shows: 하다 (to do), 가다 (to go), 되다 (to become), 알다 (to know), 받다 (to receive), 주다 (to give), 그러다 (to be like), 가지다 (to have), 떠나다 (to leave), 죽다 (to die), 먹다 (to eat), 살다 (to live), 맞다 (to be correct), 모르다 (to not know), 나가다 (to go out), 없어지다 (to disappear), 내다 (to submit), 타다 (to ride), 놓다 (to put/drop), 보내다 (to send), 앉다 (to sit), 인정하다 (to acknowledge), 놀다 (to play), 나다 (to be born), 나서다 (to come forth), 막다 (to block), 경애하다 (to respect), 계시다 (to exist), 잘못하다 (to do wrong), 넘겨주다 (to hand over)

Among these, only nine verbs overlapped between the two top-30 lists, representing an overlap rate of about 30%. This reveals a clear lexical divergence that is not visible when looking only at total type or token counts. When the comparison was expanded to the entire verb lists, 157 verbs were shared across both genres, representing 36.7% overlap.

This demonstrates that the sets of verbs used in news and comedy talk shows differ significantly. For example, the verb “경애하다 (to revere, adore a leader)” ranked second in frequency in the news corpus but only 27th in comedy talk shows. This contrast illustrates the political and

ideological nature of NK news discourse compared to the more informal, apolitical tone of comedy talk shows.

Adjectives

A pattern found in verbs is also evident in adjectives. Interestingly, as with verbs, the type count was higher in news (165 > 136), while the token frequency was higher in comedy talk shows (575 > 520). In other words, comedy talk shows used fewer types of adjectives but repeated them more often, whereas news employed a wider variety of adjectives with less repetition. This suggests that the comedy talk show corpus, reflecting the nature of everyday spoken conversation, relies heavily on a limited set of frequently used adjectives. In contrast, news discourse uses a more diverse range of adjectives to describe events and evaluations with precision. Examination of the 30 most frequently used adjectives in both genres further highlights these stylistic and functional contrasts. Among the top 30 adjectives, only six adjectives, 있다 (to be existing), 좋다 (to be good), 크다 (to be big), 고맙다 (to be thankful), 어렵다 (to be difficult), and 훌륭하다 (to be excellent)**,** appeared in both lists, accounting for just about 20% overlap**.** When the entire adjective lists were compared, only 35 adjectives overlapped, resulting in a 25.7% overlap rate**.** This relatively low overlap demonstrates that the two genres draw from largely distinct adjective sets, reflecting their different communicative purposes and stylistic norms.

In the comedy talk show corpus, adjectives such as 고맙다 (to be thankful), 곱다 (to be graceful), 안녕하다 (to be peaceful), 시원하다 (to be refreshing), 아깝다 (to be regretful), 쓸데없다 (to be useless), and 희한하다 (to be strange) frequently appeared, representing emotion-driven or personally evaluative expressions common in casual conversation. These adjectives serve to convey affect, empathy, and speaker attitude, key features of informal, spoken interaction.

By contrast, the news corpus demonstrates a strong ideological and evaluative orientation. Frequently used adjectives such as 위대하다 (to be great), 훌륭하다 (to be excellent), 자애롭다 (to be benevolent), 강력하다 (to be powerful), and 귀중하다 (to be precious) function as politically charged rhetoric, glorifying leadership, national achievement, and socialist values. Such adjectives serve not merely as descriptors but as ideological tools to reinforce loyalty and national pride.

Overall, the usage of adjectives in the two genres illustrates a clear distinction between affective and interpersonal discourse in comedy and ideological and institutional discourse in news.

Adverbs

Among the four major parts of speech (nouns, verbs, adjectives, and adverbs), adverbs showed the most distinctive contrast between the two genres.

Table 4

Token and Type Counts of Adverbs in NK News and Comedy Talk Shows

Part of Speech	Token (Frequency)		Type (Frequency)	
	News	Comedy Talk Shows	News	Comedy Talk Shows
Adverb	440 (3.05%)	907 (5.83%)	128 (5.02%)	211 (8.66%)
Conjunctive adverb	20 (0.14%)	83 (0.53%)	4 (0.16%)	9 (0.37%)

In Table 4, both token and type counts of adverbs were consistently higher in the comedy corpus. The frequency of adverbs in comedy (907, 5.83%) was almost double that in news (440, 3.05%). This can be attributed to the frequent repetition of emotional, attitudinal, and degree adverbs, such as 정말 (really), 잘 (well), 많이 (a lot), and 좀 (a bit), which reflect the speaker's feelings and evaluations. Furthermore, adverbs often co-occur in clusters within the same utterance (e.g., 얼마나 많이, 얼마나 잘), contributing to higher token frequency.

Presented below are the 30 most frequent adverbs in the News and Comedy Talk Show corpora. The adverbs are ranked by frequency within each corpus, and shared adverbs are marked in red.

News: 더 (more), 높이 (highly), 다 (all), 정말 (really), 많이 (a lot), 더욱 (further), 또 (again), 지금 (now), 얼마나 (how much), 잘 (well), 깊이 (deeply), 함께 (together), 아직 (yet), 가장 (the most), 철저히 (thoroughly), 없이 (without), 언제나 (always), 제일 (the first), 바로 (right away), 보다 (than), 특히 (particularly), 이제 (now), 대단히 (greatly), 같이 (with), 계속 (consistently), 정성껏 (carefully), 끝없이 (endlessly), 새로 (newly), 하나하나 (one by one), 또한 (again)

Comedy Talk Show: 다 (all), 또 (again), 안 (not), 좀 (little), 왜 (why), 정말 (really), 못 (not), 빨리 (fast), 얼마나 (how much), 잘 (well), 더 (more), 지금 (now), 이제 (now), 아니 (no), 딱 (perfectly), 절대로 (never), 얼른 (quickly), 그저 (just), 바로 (right away), 없이 (without), 가만(still), 모두(all together), 사실 (in fact), 거저 (for free), 도록 (so that), 몽땅 (all; colloquial), 제발 (please), 다시 (once again), 너무 (too, very), 오직 (only)

Only ten adverbs (33%) overlapped between the two top 30 lists. When expanded to the entire adverb lists, 60 adverbs overlapped, corresponding to a 28.4% overlap rate, meaning that roughly 70% of adverbs used in one genre did not appear in the other. This finding underscores the strong register-based divergence between the adverb usages in formal news and informal comedy talk shows.

A notable pattern was observed in the use of the negative adverbs "안 (do not)" and "못 (cannot)." In the comedy talk show corpus, 안 (do not) ranked third and 못 (cannot) seventh in frequency,

whereas in the news corpus, 안 (do not) appeared only at rank 69, and 못 (cannot) did not appear at all. This absence indicates that negative expressions are rarely used in news discourse, which favors assertive, positive, and ideologically aligned statements.

Conjunctive adverbs, those connecting clauses or sentences while indicating logical relations such as cause and effect, contrast, addition, or transition, also showed significant genre-specific variation. In the comedy talk show corpus, nine types of conjunctive adverbs were identified: for cause-effect 그래서, 그러니까; for contrast 그런데, 근데, 하지만; for addition 그리고; and for transition or condition 그럼, 그러면, 하긴. These appeared with a relatively high frequency of 83 tokens (0.53%), reflecting the genre's conversational and interactive style.

In contrast, the news corpus contained only four types of conjunctive adverbs (그리고, 하지만, 그러나, 그래서) with a total of 20 tokens, roughly one-fourth of the frequency observed in comedy talk shows. The prevalence of colloquial forms such as 그럼 (therefore), 그러니까 (so), and 근데 (but, by the way) in comedy talk shows demonstrates their function in structuring spoken discourse, often used toward the end of an utterance to clarify or wrap up the message.

In summary, adverbs in the comedy talk shows are both more diverse, as reflected in the higher type count, and more frequent, as shown by the higher token count, than in the news. Comedy talk shows employ a wide range of colloquial and affective adverbs repetitively, enhancing rhythm, tone, and emotional engagement. Consistent with the patterns observed in nouns, verbs, and adjectives, the roughly 30% overlap rate between the two genres confirms that they draw on distinct lexical systems, each reflecting its own communicative purpose and stylistic register.

The relatively low overlap in verb, adjective, and adverb usage between the two corpora is noteworthy. The limited shared verbs, adjectives, and adverbs suggest that informal texts introduce many additional words that learners need for broader communicative ability. If instruction is based mostly on formal materials, students may develop vocabulary that works well in formal contexts but is less useful in everyday communication. Including informal materials in instruction can therefore help expand students' vocabulary and strengthen their overall proficiency.

Other POS

Table 5 provides an overview of the distribution of sentence endings, interjections, and vocative particles across the two corpora.

Table 5

Token and Type Counts of Sentence-Final Endings, Interjections, and Vocative Particles in NK News and Comedy Talk Shows

Part of Speech	Token (Frequency)		Type (Frequency)	
	News	Comedy Talk Shows	News	Comedy Talk Shows
Sentence-final ending	317 (2.20%)	1,231 (7.91%)	5 (0.20%)	116 (4.76%)
Interjection	None	746 (4.79%)	None	103 (4.23%)
Vocative particle	None	7 (0.04%)	None	2 (0.08%)

Sentence-Final Endings

Sentence-final endings show a striking contrast between the two genres. In the news corpus, only five formal sentence-final endings were identified, whereas the comedy talk show corpus contained as many as 116 distinct forms, including a wide range of colloquial endings. Their token frequency in the comedy talk show was also substantially higher.

- Endings used in the news (5 types): 습니다, ㅂ니다, 다, 습니까, ㄴ답니다
- Colloquial endings used in comedy talk shows (20 types): 구먼요, 다니요, 는구먼, ㄴ대, 대, ㄴ답니다, ㄴ데요, 던데, 라우, 아야지, 는지, 로군, 외다, 라요, 로구먼, 는군요, 더라, 든가, 너라, 라니

This contrast reflects the stylistic nature of NK news, which maintains a formal, standardized, and monotonous tone aimed at delivering information or official statements. In contrast, the comedy talk show genre exhibits rich variation in colloquial endings, used for declarative, interrogative, imperative, exclamatory, and emphatic expressions, that convey the speaker's emotions, attitudes, intimacy, and social dynamics. The diversity of these sentence-final endings underscores that comedy talk shows are highly interactive and emotionally expressive spoken discourse, while news remains formal and non-interactive.

Interjections

Interjections offer another clear point of contrast between the two genres. No interjections appeared in the news corpus, whereas the comedy talk show corpus included over 100 distinct types and 746 tokens. Interjections typically signal spontaneous emotional responses such as surprise, frustration, admiration, or sympathy and help establish rapport by creating an immediate emotional connection between speakers and listeners. In contrast, news discourse maintains an emotion-neutral, objective style that leaves little room for such expressions. As a result, the presence or absence of interjections provides a clear indication of the emotional stance and communicative distance characteristic of each genre.

- Interjections used in the news (0): None
- Interjections used in comedy talk shows (20 examples): to signal hesitation and uncertainty (뭐, 어, 자, 글쎄, 저); to express sympathy (아이고, 에그, 참, 에고, 에구, 아유, 아이구); for displeasure and frustration (흥, 에이, 아차, 에라); to show a recognition or agreement (그래, 아하)

Vocative Particles

Vocative particles also appeared only in the comedy talk show corpus and were completely absent in news. These particles, such as 야 and 아, do not carry specific lexical meaning; rather they are used to directly address a person or audience to draw attention, serving a conversational and interactive function. The word 동무 (comrade) is a lexical item specific to NK usage and serves as a vocative form of address. Their presence reflects the dialogic and participatory structure of NK comedy talk shows, where emotional exchange and interpersonal connection play a central role. In contrast, NK news delivers information in a one-way, non-interactive format directed at a mass audience, leaving no place for vocative markers. Thus, the appearance of vocative particles exclusively in comedy talk show highlights the oral, interpersonal, and audience-engaging nature of spoken discourse, distinguishing it from the formal, impersonal style of news reporting.

Limitations

While conducting the morphological analysis, some NK endings were manually adjusted to their SK equivalents because the morphological analyzer used in this study was originally designed for the SK language and did not recognize certain NK forms. Therefore, this adjustment may have influenced the distributional results to a limited extent. However, a recently upgraded morphological analyzer (https://kcorpus.korean.go.kr/), specifically designed to accommodate NK linguistic features, has been developed by the National Institute of Korean Language (NIKL) in South Korea. Future research employing improved NK morphological analysis tools is expected to achieve greater accuracy and reduce the need for manual normalization.

CONCLUSION

This study examined the lexical differences between NK formal speech as represented in news broadcasts and informal speech as found in comedy talk shows through a corpus-based POS analysis. A comparison of type and token distributions across major parts of speech revealed clear distinctions between the two genres in NK.

Table 6

Comparison of Token and Type Counts in NK News and Comedy Talk Show Data

Part of Speech	Token	Type
Common Nouns	News > Talk Show	News > Talk Show
Verbs	News < Talk Show	News > Talk Show
Adjectives	News < Talk Show	News > Talk Show
Adverbs	News < Talk Show	News < Talk Show

The findings of this study can be interpreted meaningfully within Bachman and Palmer's (2010) TLU domain framework. As the analysis demonstrates, the lexical profiles of NK news broadcasts and NK comedy talk shows differ systematically in POS distribution, type-token ratios, and lexical overlap. These differences are not merely stylistic; rather, they reflect distinct communicative purposes, discourse conventions, and interactional conditions. In TLU terms, the two genres represent separate domains characterized by different task demands, participant roles, and pragmatic expectations. The higher proportion of common nouns in news discourse corresponds to its expository, information-dense function within a formal setting, while the greater frequency of verbs, adjectives, adverbs, and repetition in comedy talk shows reflects the interactive, affective, and situationally embedded nature of informal spoken communication. Thus, the observed lexical variation aligns with the TLU principle that linguistic features are shaped by the contextual parameters of language use.

The lexical overlap analysis further reinforced the differences between the two genres as shown in Table 7.

Table 7

Lexical Overlap Percentage of the Two Genres by POS

Part of speech	Common Noun	Verb	Adjective	Adverb
Number of Overlapping Items	254	154	35	60
Lexical Overlap Ratio (%)	27.1%	36.7%	25.7%	28.4%

From a domain representativeness perspective, the limited lexical overlap—approximately 30% across major parts of speech—indicates that reliance on a single register provides access to only a partial segment of the broader NK lexicon. If instructional materials are drawn predominantly from formal news discourse, learners are effectively trained within a restricted TLU domain. While this focus may support proficiency in formal or informational contexts, it does not necessarily equip learners to interpret language in informal, interactional settings. Bachman and Palmer (2010) emphasize that meaningful language development depends on alignment between instructional input and the range of real-world communicative situations learners must handle. The present findings suggest that current NK instruction may underrepresent the lexical resources associated with conversational and emotionally expressive discourse, thereby limiting learners' register flexibility. The implication is not that exposure to informal media directly determines performance on assessments such as Defense Language Proficiency Test 5 (DLPT5), but rather that reliance on formal materials alone provides access to only a limited segment of

the broader NK lexicon. When instruction concentrates primarily on formal registers, learners may develop strong comprehension within that domain while remaining less prepared for communicative contexts that depend on different high-frequency vocabulary.

Although this study clearly identified lexical differences between the news and comedy talk show corpora, the disparity in raw corpus length presents a methodological limitation that may have influenced the overlap analysis. Because each corpus varies in size, the absolute number of shared lexical items must be interpreted with caution. While the POS percentage distributions remain stable, as each reflects proportions calculated within its respective corpus, the measures of lexical overlap are based on raw counts and are therefore inherently more sensitive to differences in corpus length. To mitigate this effect, overlap indices were interpreted in relative rather than absolute terms, and corpus size was taken into account when comparing patterns across registers. Accordingly, differences in lexical overlap should be understood as indicative of distributional tendencies rather than direct magnitude comparisons. Future research would benefit from balanced corpora to allow more precise comparisons.

Future research should extend this line of inquiry by examining the relationship between expanded TLU domain coverage and measurable gains in language proficiency. In particular, studies could investigate whether systematic exposure to vocabulary drawn from both formal and informal NK domains correlates with improved comprehension, production, and register flexibility. Such research would help determine the extent to which broader lexical exposure translates into functional communicative competence, thereby providing further empirical grounding for corpus-informed, TLU-aligned curriculum development.

Authors

Mi Hye Lee, Ph.D., is a Korean language instructor at DLIFLC with extensive language teaching experience in both South Korea and the United States. Her professional interests include curriculum development, instructional materials, and assessment practices in language education.

Myoyoung Kim, Ph.D., currently serves at DLIFLC as an educational measurement specialist and institutional evaluator. Her career in language education spans teaching, curriculum development, teacher training, and assessment. With a background in linguistics, she has a strong interest in data management and analysis and enjoys using data-driven approaches in her work.

REFERENCES

Bachman, L. F., & Palmer, A. S. (2010). Language assessment in practice: Developing language assessments and justifying their use in the real world. Oxford University Press.

Biber, D. (2001). On the complexity of discourse complexity: a multi-dimensional analysis. In S. Conrad & D. Biber (Eds.), *Variation in English: Multi-dimensional studies* (pp. 177–202). Longman.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25(3)*, 371–405.

Baron, N.S. (2013). *Words Onscreen: The Fate of Reading in a Digital World*. Oxford University Press.

Cheong, Y. (2021). A study on the concept and classification criteria of North Korean. *Baedalmal,* 69, 213–241. https://doi.org/10.52636/KL.69.7

Curry, N., & Mark, G. (2024). Using corpus linguistics in materials development and teacher education. *Second Language Teacher Education,2(2)*, 187–208. doi:10.1558/slte.25727

Eum, I., Seo, H., & Kwon, S. (2021). Development of North Korean vocabulary textbooks for South Koreans*. Journal of CheongRam Korean Language Education,* 80, 413–459. https://doi.org/10.26589/jockle..80.202103.413

Kim, J. (2015). A critical review on education for understanding of North Korean language in Korean education*. Korean Language Education Research*, 58, 143–170.

Kim, N. (2024). A study on improving Korean language education for unification: Focusing on teaching North Korean as a regional dialect (Thesis). Yonsei University.

Kim, Y. J., & Biber, D. (1994). A Corpus-Based Analysis of Register. *Sociolinguistic perspectives on register*, 157.

Latham, E. (2025). Using corpus-driven TLU domain analysis to increase the authenticity of listening passages. *International Journal of English for Academic Purposes: Research and Practice, vol 5*, n 2. https://doi.org/10.3828/ijeap.2025.8

Li, D., Noordin, N., Ismail, L., & Cao, D. (2025). A Systematic Review of Corpus-Based Instruction in EFL Classroom. *Heliyon 11(2)*, 10.1016/j.heliyon.2025.e42016

Matang, P., & Narathakoon, A. (2025). A Corpus-Based Study of Lexical Bundles of Keywords Found in Online News Articles, *Thai TESOL Journal vol. 38, issue 1*.

Moser, J. (2020) Evaluating Arabic textbooks: A corpus-based lexical frequency study, *International Journal of Applied Linguistics, vol.31,* issue 2. https://doi.org/10.1111/ijal.12321

Oh, J. (2015). *A study on education methods of North Korean vocabularies* (Thesis). Graduate School of Education, Kyung Hee University.

Reppen, R. (2001). Register variation in student and adult speech and writing. In S. Conrad & D. Biber (Eds.), *Variation in English: Multi-dimensional studies* (pp. 187–199). Longman.

Tannen, D. (1985). Relative focus on involvement in oral and written discourse. In D. R. Olson, N. Torrance, & A. Hildyard (Eds.), *Literacy, language, and learning: The nature and consequences of reading and writing* (pp. 124–147). Cambridge University Press.

Yeo, E. (2021). *Several issues on the morphological analysis in North Korean corpus* (Thesis). Jeonbuk University.

APPENDIX A

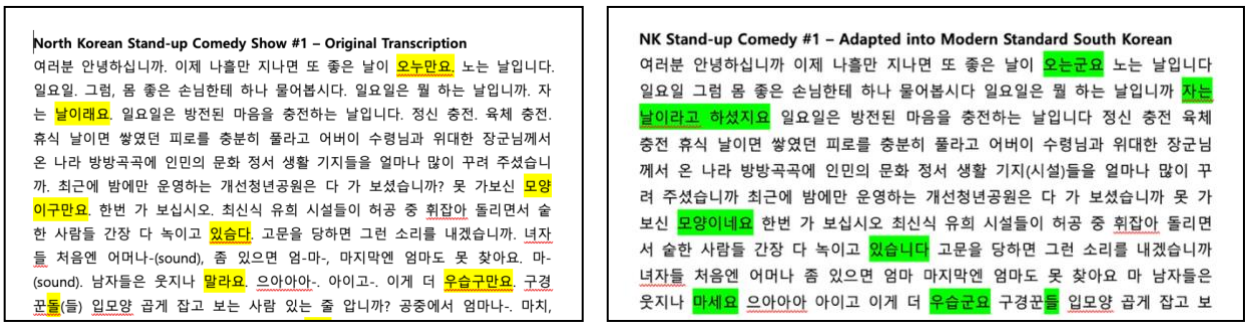
Video Information

	Topic	Genre	Length (min:sec)	Production Year	Link
1	일요일 (Sunday)	Solo comedy Talk show	10:05	Unknown	https://www.youtube.com/watch?v=2-GSHwx6HY0
2	미신을 믿지 말자 (Let's not believe in superstitions)	Two-person comic play	10:40	Unknown	https://www.youtube.com/watch?v=dasJfVO8krE
3	단호한 결심 (A firm resolution)	NK *Jaedam* (comic performance)	15:27	1986	https://www.youtube.com/watch?v=WaY5aCoxCgM&list=PLMHnawEno0Eqk40HcGQjcO2FxLwW_DWut&index=1
4	인심 좋은 작업반장 (A kind-hearted work team leader)	Comedy sketch	10:54	1987	https://www.youtube.com/watch?v=DTTjeCRBWBk
5	정 (Warm Affection)	Comedy sketch	12:00* (20:27)	2016	https://www.youtube.com/watch?v=x02wfE_RQUw&list=PLMHnawEno0Eqk40HcGQjcO2FxLwW_DWut&index=4
6	오가는 정 (Affection Given and Returned)	Comedy sketch	20:00* (29:35)	Unknown	https://www.youtube.com/watch?v=Vy5EpgW2IeI (*currently unavailable)
7	떠나는 마음 보내는 마음 (A Heart Leaving, A Heart Letting Go)	Comedy sketch	20:00* (30:47)	2005	https://www.youtube.com/watch?v=_D4l4ZQIwck

*For videos #5~#7, only the first 12 or 20 minutes (out of the total length shown in parentheses for each video) were transcribed to avoid lexical concentration on a single topic.

APPENDIX B

Example of transcript verification



North Korean Stand-up Comedy Show #1 – Original Transcription

여러분 안녕하십니까. 이제 나흘만 지나면 또 좋은 날이 오누만요. 노는 날입니다. 일요일. 그럼, 몸 좋은 손님한테 하나 물어봅시다. 일요일은 뭘 하는 날입니까. 자는 날이래요. 일요일은 방전된 마음을 충전하는 날입니다. 정신 충전. 육체 충전. 휴식 날이면 쌓였던 피로를 충분히 풀라고 어버이 수령님과 위대한 장군님께서 온 나라 방방곡곡에 인민의 문화 정서 생활 기지들을 얼마나 많이 꾸려 주셨습니까. 최근에 밤에만 운영하는 개선청년공원은 다 가 보셨습니까? 못 가보신 모양이구만요. 한번 가 보십시오. 최신식 유희 시설들이 허공 중 휘잡아 돌리면서 술한 사람들 간장 다 녹이고 있습다. 고문을 당하면 그런 소리를 내겠습니까. 녀자들 처음엔 어머나-(sound), 좀 있으면 엄-마-, 마지막엔 엄마도 못 찾아요. 마-(sound). 남자들은 웃지나 말라요. 으아아아-. 아이고-. 이게 더 우습구만요. 구경꾼들(들) 입모양 곱게 잡고 보는 사람 있는 줄 압니까? 공중에서 엄마나-. 마치,

NK Stand-up Comedy #1 – Adapted into Modern Standard South Korean

여러분 안녕하십니까 이제 나흘만 지나면 또 좋은 날이 오는군요 노는 날입니다 일요일 그럼 몸 좋은 손님한테 하나 물어봅시다 일요일은 뭘 하는 날입니까 자는 날이라고 하셨지요 일요일은 방전된 마음을 충전하는 날입니다 정신 충전 육체 충전 휴식 날이면 쌓였던 피로를 충분히 풀라고 어버이 수령님과 위대한 장군님께서 온 나라 방방곡곡에 인민의 문화 정서 생활 기지(시설)들을 얼마나 많이 꾸려 주셨습니까 최근에 밤에만 운영하는 개선청년공원은 다 가 보셨습니까 못 가보신 모양이네요 한번 가 보십시오 최신식 유희 시설들이 허공 중 휘잡아 돌리면서 술한 사람들 간장 다 녹이고 있습니다 고문을 당하면 그런 소리를 내겠습니까 녀자들 처음엔 어머나 좀 있으면 엄마 마지막엔 엄마도 못 찾아요 마 남자들은 웃지나 마세요 으아아아 아이고 이게 더 우습군요 구경꾼들 입모양 곱게 잡고 보

*The left side shows the output before manual verification, and the right side presents the modified version after verification. Different colors were used to indicate how spellings of the same words were changed. On the right side, green highlight instances in which the same lexical item appears with modified spellings; the corresponding words on the left side were originally transcribed. The colors do not signal differences in meaning, but rather variation in transcription certainty and orthographic representation of the same word. This procedure was applied to NK news broadcasts to minimize the error rate.

APPENDIX C

POS Distribution and Full Results for the News and Comedy Corpora

C-1. Overview

Part of Speech	Tag Set	News	Comedy Talk Show
Common Noun	NNG	3,821 (26.51%)	2,790 (17.92%)
Verb	VV	1,586 (11%)	1,940 (12.46%)
Adjective	VA	520 (3.61%)	575 (3.69%)
General Adverb	MAG	440 (3.05%)	907 (5.83%)
Others	EF, EC, IC, NNP, NNB, NP, MM, etc.	463 (18.15%)	985 (29.8%)
Total		14,415	15,567

C-2. Comparison of POS count: Tokens

	Part of Speech	Tag Set	News	Comedy Talk Shows
1	Common noun	NNG	3,821 (26.51%)	2,790 (17.92%)
2	Proper noun	NNP	247 (1.71%)	115 (0.74%)
3	Adnominal particle	JKG	499 (3.46%)	128 (0.82%)
4	Adnominal modifier	MM	214 (1.48%)	329 (2.11%)
5	Adnominal ending	ETM	1,026 (7.12%)	600 (3.85%)
6	Positive copula	VCP	144 (1.00%)	352 (2.26%)
7	Pronoun	NP	282 (1.96%)	733 (4.71%)
8	Verb	VV	1,586 (11.00%)	1,940 (12.46%)
9	Verb-derivational suffix	XSV	14 (0.10%)	18 (0.12%)
10	Noun-derivational suffix	XSN	491 (3.41%)	259 (1.66%)
11	Nominalizing ending	ETN	74 (0.51%)	34 (0.22%)
12	Objective particle	JKO	695 (4.82%)	414 (2.66%)
13	Complement particle	JCM	29 (0.20%)	38 (0.24%)
14	Auxiliary verb/adjective	VX	377 (2.62%)	302 (1.94%)
15	Auxiliary particle	JX	394 (2.73%)	599 (3.85%)
16	Adverbial particle	JKB	664 (4.61%)	393 (2.52%)
17	Negative copula	VCN	12 (0.08%)	44 (0.28%)
18	Prefinal ending	EP	360 (2.50%)	490 (3.15%)
19	Numeral	NR	36 (0.25%)	88 (0.50%)
20	Number	SN	93 (0.65%)	8 (0.05%)
21	Connective ending	EC	1,240 (8.60%)	1,383 (8.88%)
22	Dependent noun	NNB	293 (2.03%)	408 (2.62%)
23	Quotative particle	JKQ	3 (0.02%)	5 (0.03%)
24	Adverb	MAG	440 (3.05%)	907 (5.83%)
25	Conjunctive adverb	MAJ	20 (0.14%)	83 (0.53%)
26	Conjunctive particle	JC	159 (1.10%)	55 (0.35%)
27	Final ending	EF	317 (2.20%)	1,231 (7.91%)
28	Subject particle	JKS	356 (2.47%)	492 (3.16%)
29	Nominal prefix	XPN	8 (0.06%)	10 (0.06%)
30	Adjective	VA	520 (3.61%)	575 (3.69%)
31	Adj.-derivational suffix	XSA	1 (0.01%)	1 (0.01%)
32	Interjection	IC	None	746 (4.79%)
33	Vocative particle	JKV	None	7 (0.04%)
	Total		14,415	15,567

C-3. Comparison of POS counts: Types

	Part of Speech	Tag Set	News	Comedy Talk Shows
1	Common noun	NNG	1,308 (51.31%)	937 (38.45%)
2	Proper noun	NNP	98 (3.84%)	71 (2.91%)
3	Adnominal particle	JKG	1 (0.04%)	1 (0.04%)
4	Adnominal modifier	MM	27 (1.06%)	37 (1.52%)
5	Adnominal ending	ETM	10 (0.39%)	17 (0.70%)
6	Positive copula	VCP	1 (0.04%)	1 (0.04%)
7	Pronoun	NP	18 (0.71%)	44 (1.81%)
8	Verb	VV	487 (19.11%)	428 (17.56%)
9	Verb-derivational suffix	XSV	4 (0.16%)	3 (0.12%)
10	Noun-derivational suffix	XSN	31 (1.22%)	20 (0.82%)
11	Nominalizing ending	ETN	3 (0.12%)	2 (0.08%)
12	Objective particle	JKO	3 (0.12%)	3 (0.12%)
13	Complement particle	JCM	2 (0.08%)	2 (0.08%)
14	Auxiliary verb/adjective	VX	20 (0.78%)	22 (0.90%)
15	Auxiliary particle	JX	17 (0.67%)	28 (1.15%)
16	Adverbial particle	JKB	18 (0.71%)	16 (0.66%)
17	Negative copula	VCN	1 (0.04%)	1 (0.04%)
18	Prefinal ending	EP	7 (0.27%)	8 (0.33%)
19	Numeral	NR	15 (0.59%)	22 (0.90%)
20	Number	SN	39 (1.53%)	7 (0.29%)
21	Connective ending	EC	67 (2.63%)	107 (4.39%)
22	Dependent noun	NNB	56 (2.20%)	62 (2.54%)
23	Quotative particle	JKQ	1 (0.04%)	2 (0.08%)
24	Adverb	MAG	128 (5.02%)	211 (8.66%)
25	Conjunctive adverb	MAJ	4 (0.16%)	9 (0.37%)
26	Conjunctive particle	JC	5 (0.20%)	8 (0.33%)
27	Final ending	EF	5 (0.20%)	116 (4.76%)
28	Subject particle	JKS	4 (0.16%)	3 (0.12%)
29	Nominal prefix	XPN	3 (0.12%)	7 (0.29%)
30	Adjective	VA	165 (6.47%)	136 (5.58%)
31	Adj.-derivational suffix	XSA	1 (0.04%)	1 (0.04%)
32	Interjection	IC	None	103 (4.23%)
33	Vocative particle	JKV	None	2 (0.08%)
	Total		2,549	2,437