

**Nikolay STANKOV**

**St. Cyril and St. Methodius University of Veliko Tarnovo, Bulgaria**

## **DEVELOPMENT TRENDS IN THE MORPHOLOGICAL-TYOLOGICAL PROFILE OF GERMAN**

*The article presents the results of an analysis of samples extracted from a total of six corpora containing literary, scientific and journalistic texts stemming from fixed periods within the 20th and 21st centuries. The aim is to discover whether and to what extent the morphology of the German language has become more or less analytic or synthetic and what quantitative morphological-typological differences there are between the three major text types: literary, scientific, and journalistic. The concept of grammaticity is also briefly discussed and applied to the analysis. The study finds, among other things, that 21st-century scientific texts tend to exhibit a higher level of syntheticity than the other two text types, that German has simultaneously become less analytic and less synthetic over the past 100 years, and that, as a result, it encodes less grammatical information nowadays than it did at the outset of the 1920s.*

**Keywords:** *typological profiling; analyticity index; syntheticity index; grammaticity; German; corpus samples.*

### **1. Introduction**

The first two decades of the 21st century have been a time of unprecedented progress and growth in the field of information and computer technology. The rapid advancement of the technical capabilities of computers and their increasing accessibility and widespread use in various spheres of everyday life has allowed humanity to swiftly accumulate vast amounts of data and analyze them quickly, precisely and efficiently. The novel technologies are well applied and accepted in both science and humanities, fuelling empirical, data-driven research across the globe. As a discipline that allows for a lot of cross-disciplinary collaboration, linguistics has not remained isolated from these developments.

Linguistics is a multifaceted discipline with many branches and subbranches. Consequently, the way information and computer technologies are implemented in linguistic studies may vary substantially. However, one significant advantage of the new technological developments that may apply to nearly all areas of linguistic research is the fact that the ease and swiftness of data analysis provided by modern computers allows for empirical, quantitative answers to questions that could

previously only be answered in a qualitative and often more subjective way. Of course, the merits of a qualitative study and the resulting conclusions should not be understated, yet data-driven quantitative research is a powerful tool for testing and solidifying these conclusions.

While the new data-driven approaches can make valuable contributions towards answering the newer questions in linguistics, e.g. those involving the utilization of natural language by machines and the development of artificial intelligence, the discussion of questions that have characterized linguistic discourse for centuries can also benefit significantly. The discussion that this study aims to revisit involves the issue of the diachronic morphological-typological development of Indo-European languages. The terms *analytic* and *synthetic* with regard to languages have been in use for nearly two centuries and are some of the basic typological concepts that linguistics students encounter in their first years on campus. Yet, in spite of the aforementioned recent technological progress, the amount of quantitative research involving this typological framework remains fairly low, even though the analytic-synthetic distinction, which is often construed as a continuum, remains a frequently used instrument for typological classification and has been the topic of many discussions in academia.

## **2. Aims and Methodology of the Study**

### **2.1. Aims**

The current study attempts to revisit the topic of analyticity and syntheticity in Indo-European languages from a more data-driven quantitative perspective, taking into consideration both the synchronic and diachronic level of language study. The Indo-European language family includes a large number of individual languages and dialects, which is why the focus now is only on German and on its typological development. However, the involvement of other languages from the family and a comparison between them are planned for future publications. The main aim here is to determine the direction of the diachronic morphological-typological development of the German language, roughly within the framework of the 20th century, and to provide quantitative evidence for it. Another goal is to observe the quantitative morphological-typological differences between individual text types (literary, scientific and journalistic) in the German language, and possibly to discover how these have changed over the past 100 years. Attention is also drawn to the concept of grammaticity (Szmrecsányi 2009), which represents the total amount of grammatical information expressed via free or bound grammatical markers in a text, and what differences can be observed in this metric across the investigated corpora.

### **2.2. Corpus Composition and Method of Analysis**

The methodology of this study is based on Szmrecsányi and Kortmann (2011), Szmrecsányi (2016) and Horsch (2021), which are themselves inspired by Green-

berg (1960), but some notable alterations were introduced to their approach in order to achieve a more detailed and precise quantitative morphological-typological profile of the language and the investigated text types. Six corpora of approx. 50,000 word tokens each were initially compiled to serve as sources for the extraction of samples, which were to be analyzed, following the method described below. Three of these corpora contained texts from the period 1918–1922, and the other three from the period 2018–2022. Within each of the two groups of three corpora, one corpus contained literary texts (excerpts from novels by German authors, published within the respective time frame), one scientific texts (excerpts from academic publications by German authors, published within the respective time frame) and one journalistic texts (articles and reports published in German newspapers or online media within the respective time period). Five 1,000-word token samples were extracted from each of these six corpora. They were selected randomly, using a random number generator. The random process selected entire sections of coherent text from each corpus, which were to be extracted as samples, and not individual words. This is important because Szmrecsányi and Kortmann (2011) selected every word in their samples randomly, resulting in what could only be incoherent sample texts, while in this study, it was aimed to obtain samples containing coherent text.

Once selected and extracted, the samples were fed into the POS tagger software TagAnt, developed by Lawrence Anthony (2022). The software employed the tagset used in the text annotation tool TreeTagger<sup>1</sup> by Schmid (1994; 1995). TagAnt attached a tag to each individual token in the sample, depending on its part-of-speech class and grammatical function within the sentence. The tagged samples were then manually examined and the elements that bore grammatical meaning were allocated either to the group of analytic tokens or to the group of synthetic ones. The following parts of speech were marked as analytic tokens: prepositions, circumpositions, postpositions, articles (both definite and indefinite), conjunctions, subjunctives, pronouns (including personal, demonstrative, possessive, reflexive, interrogative, relative, and indefinite pronouns), pronominal adverbs, the negative particle *nicht* ‘not,’ the particle *zu* ‘to,’ modal verbs (including the verb *lassen* ‘let’ when it is not used as a full verb), auxiliary verbs and existential uses of the verb *geben* ‘give’ (as part of the phrase *es gibt* ‘there is’). Once a sample was processed, the number of analytic tokens was determined. The number of analytic tokens within a sample is its analyticity index (AI). A mean analyticity index was calculated for each corpus by adding up the AI of every sample obtained from the corpus and dividing the result by the number of samples.

In addition to the AI, a syntheticity index (SI) was calculated for every sample. It is equal to the number of syntheticity tokens in the sample. In contrast to the analyticity tokens, which are entire word tokens, syntheticity tokens are morphemes. During the manual processing and analysis of the samples, individual mor-

<sup>1</sup> The German tagset file is available at <https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>.

phemes that carried grammatical meaning within the boundaries of word tokens were identified and marked. The following types of morphemes were counted as synthetic tokens: case, number and grammatical gender markings on nouns, pronouns, articles and adjectives (e.g. *-en*, *-em*, *-es*, etc.), including vowel mutation or umlaut (e.g. the vowel *ä* in *Väter* ‘fathers,’ a form derived from *Vater* ‘father’<sup>2</sup>); verb conjugation endings (e.g. *-e*, *-t*, *-en*); tense and participle markings on verbs (e.g. *-te*, *ge-*, *-t*, *-d*), including ablaut (e.g. the root vowel *a* in *bat* ‘asked,’ which is derived from the present form *bitten* ‘ask’); subjunctive markings on verbs (e.g. the vowel *ü* in *würde* ‘would,’ as derived from *wurde* ‘became’); conflated forms consisting of preposition and article (e.g. *im* ‘in the’ from *in* ‘in’ and *dem* ‘the’); comparative and superlative markings on adjectives (e.g. *-er* and *-st*), including vowel mutation (e.g. the vowel *ä* in *kälter* ‘colder,’ as derived from *kalt* ‘cold’). After the processing was completed, the number of synthetic tokens in each sample was counted<sup>3</sup> and the result was considered the syntheticity index (SI) of the respective sample. A mean syntheticity index was calculated for each corpus by adding up the SI of every sample obtained from the corpus and dividing the result by the number of samples. A grammaticity index (GI) was calculated by adding together the mean AI and SI for each text type. The concept of grammaticity is further explained below. The AI and SI of the samples obtained from the corpora containing 21st-century texts, as well as the mean indices (mean AI, mean SI and mean GI), can be seen in Table 1. The same information about the samples obtained from the corpora containing 20th-century texts can be found in Table 2.

### 2.3. Corpus Size

Compared to modern corpora of the German language available online for free<sup>4</sup>, a corpus of 50,000 word tokens appears quite insignificant. A few arguments can be made to justify the choice of this corpus size over the vastly larger alternatives. Firstly, each of the corpora compiled for this study is focused on a particular text type and historical period and is, therefore, not intended to be representative of the entire spectrum of language usage, but only of a particular sphere of application (e.g. scientific discussion or journalism). Secondly, the corpora themselves are not the object of analysis. Instead, it is the samples that are extracted from the corpora that are subjected to analysis. Given that the corpus building process involved a wide

---

<sup>2</sup> Note that two morphemes serving the same grammatical function within the word were counted as one synthetic token (e.g. the mutation of *o* to *ö* and the ending *-e* in *Wölfe* ‘wolves,’ as derived from *Wolf* ‘wolf’).

<sup>3</sup> Note that a single word token could contain several morphemes, carrying individual and separate grammatical information, and, therefore, one word token could yield several synthetic tokens.

<sup>4</sup> The DWDS corpora of German number a total of more than 50 billion word tokens (<https://www.dwds.de/d/korpora>), while the Mannheim German Reference Corpus (DeReKo) offers 46.9 billion word tokens of written German (<https://www.ids-mannheim.de/en/digspra/corpus-linguistics/projects/corpus-development/>).

array of sources, the resulting corpora should provide samples that are sufficiently representative of the respective text types and historical periods, even though the size of each individual corpus is relatively small. Lastly, with online corpora there is the issue of accessibility. The large online corpora surveyed in preparation for this work did not grant unlimited access to the corpus's text, even though they offered comprehensive ways of searching their databases and exploring the resulting hits. The extraction of 1,000-word text samples from them is a difficult endeavour for the regular user, if it is possible at all, and they were therefore found to be incompatible with the current study's design. Similar arguments about the sufficiency of small homogenous corpora for language studies with specific goals are put forward by Bowker and Pearson (2002: 48, as cited in Seghiri 2014: 89) and Kock (1997: 292).

#### 2.4. Term Clarifications

Before delving into the analysis of the corpus data, a few remarks should be made about the terms used in this article. The term *text type*, as used in the study, needs to be clarified, as it could be understood in different ways, and there are various classifications of text types that have been proposed over the years. The definition adopted here follows Fischer (2009), according to whom "eine Textsorte umfasst also alle diejenigen Texte, welche die gleiche kommunikative Funktion realisieren, ähnlich strukturiert sind und in einem übereinstimmenden Kontext stehen." "A text type thus includes all those texts which fulfil the same communicative function, are similarly structured and are situated in a common context." (Fischer 2009: 16). This definition is based on Brinker (2005: 144, as cited in Fischer 2009: 16). The current study only looks at literary, scientific and journalistic texts, but it is freely acknowledged that these are fairly broad categories, and more specific and detailed classifications within each one of these categories are possible.

For the purposes of this work, the term *literary texts* encompasses only prose: novels, short stories and other narrative-based literary works. Poetry and drama, although also major forms of literature, are not included, because they often have a more varied and unique structure and form, which makes them harder to compare to the other text types in the study. Additionally, such texts were left out of the study and, consequently, of the literary corpora, in order to ensure the homogeneity of the literary corpus and to avoid inconclusive analysis results. The term *scientific texts* applies to texts focusing on particular objects or topics of scientific inquiry and analysis that are published in journals, textbooks, or other officially acknowledged publications of the scientific community. Popular science texts are excluded from the category, because they often include features characteristic of other text types, having a larger and more heterogeneous target group. *Journalistic texts* include both reporting and opinion pieces, although the two types show some differences in structure. Nevertheless, both are undeniably characteristic of journalism and journalistic publications, which is why both types of publications were included in the respective corpus.

The term *morphological-typological profile* is more rarely encountered and should also be clarified in the current context. In the study, it is used mainly in relation to the concepts of analyticity and syntheticity, which are defined below. The term comprises the level of analyticity and syntheticity of the morphological structure of a natural language system. It is also important to note here that the paper aims to examine and discuss the morphological-typological profile of German only from a quantitative perspective, since a qualitative analysis would be a much more substantial undertaking, requiring much more time and space.

For the concepts of *analyticity* and *syntheticity*, this study follows the work of Szmrecsányi (2016: 95), who defines:

– “*formal grammatical analyticity* as covering all coding strategies that convey grammatical information via free grammatical markers, which in turn are defined as synsemantic (see Marty 1908) word tokens devoid of independent lexical meaning;” and

– “*formal grammatical syntheticity* as covering all those coding strategies where grammatical information is signalled by bound grammatical markers.”

The only alteration made to the definition of analyticity is that the synsemantic nature of free grammatical markers is limited to the context in which the respective marker is used. The free grammatical marker is viewed as synsemantic in the particular context, where it is used, but is not necessarily synsemantic in all possible contexts of its use. In particular, such is the case with modal verbs and the verbs *lassen* ‘let’ and *geben* ‘give.’

Lastly, this work makes use of the concept of *grammaticity*, which is borrowed from Szmrecsányi (2009: 322), where he defines it as a notion that “comprises all explicit grammatical markers, but not word order,” and as “a ratio of the total number of grammatical markers [...] in a text to the total number of words.” As the definition implies, this metric measures the total amount of grammatical information in a text and is therefore a useful tool in the exploration of the morphological-typological profile of a language.

### **3. Corpus Data Analysis**

#### **3.1. Literary Text Type**

The analysis of the three 21st-century corpora showed some interesting differences in the morphological-typological profiles of the three investigated text types. Starting with the literary corpus, we observe a mean AI of 456 and a mean SI of 415. Compared to the results from the other two 21st-century corpora (see Table 1), this is the only corpus where the mean AI is higher than the mean SI. The increased AI may be due to the presence of dialogue in narrative-driven literary texts. While an analysis of oral speech is not included in the current study, and it cannot be foreseen what results such an analysis would yield, it may be surmised that, because of its focus on economy, familiarity and ease, everyday speech between close individuals

like friends or family members features more pronouns, prepositions and auxiliary verbs than more conservative and formal written texts. A more comprehensive part-of-speech analysis of the samples extracted from the corpora compiled for this study could show whether this is a viable hypothesis. However, such an analysis would go beyond the scope of this article.

The samples obtained from the corpus of literary texts from the 21st century have a higher mean AI than the samples from both the scientific and journalistic corpora from this period. At the same time, the mean SI of literary texts is lower than the mean SI of both scientific and journalistic texts. The lower syntheticity of the literary samples may also be due to the presence of dialogue in them, as the parts of speech that commonly occur in dialogue (pronouns, prepositions, adverbs, particles, proper names) are rarely subject to inflection. On the other hand, however, literary texts that feature more detailed descriptions of environments or actions could exhibit a higher level of syntheticity, as they are likely to contain more adjectives and verbs, both of which are subject to inflection. This may be a reason why Sample 1 from the corpus with 21st-century literary texts has a significantly higher SI than the other samples from the same corpus.

The results from the corpus with 20th-century literary texts are similar to those from the corresponding corpus with 21st-century texts. The mean AI of 457 is almost identical, while the mean SI is only slightly lower (425). We also see the same tendency for analyticity to be higher than syntheticity. This is not surprising, considering that, in the first decades of the 20th century, the major genres of literary prose and the corresponding styles of writing were already well established. Yet, a study of older literary prose texts from earlier centuries could reveal a different picture. In any case, the morphological-typological profile of literary prose has apparently changed only very slightly over the past 100 years, with the only observable trend being the one towards lesser syntheticity, which is in fact common to all three text types investigated.

In addition to the differences in AI and SI, Table 1 also shows a meaningful difference in the grammaticity index (GI) of the two corpora containing literary texts. Similar to the other two indices, the difference is comparatively small – from 882 in the 20th-century corpus to 871 in the 21st-century corpus, – but as we will see in the sections below, it is in line with observations made for the other two text types. The decline in mean GI suggests a decrease in the amount of grammatical information that the text encodes (for a discussion of the relation between GI and the overt grammatical information encoded in texts, cf. Szmrecsányi 2009), which is an intriguing development, especially if it can be attributed to the entire language system.

### **3.2. Scientific Text Type**

The analysis of the corpus comprising scientific publications from the 21st century yielded a mean AI of 426 and a mean SI of 474. Unlike the literary texts,

whose analyticity is higher than their syntheticity, the situation here is reversed. In addition, the difference between the two metrics is larger than in the literary texts. These differences could be expected, considering the significant differences between the literary and scientific writing styles. Firstly, scientific texts are unlikely to feature dialogue or narrative while focusing on detailed and precise technical descriptions and explanations. Such texts are likely to feature more complicated verb and noun phrases with more adjectives and verb participles, as well as passive constructions. In German, adjectives, verbs and nouns are the parts of speech that are most commonly subject to inflection. Again, to test the reliability of this explanation, a more detailed part-of-speech analysis of the samples needs to be performed.

One additional factor contributing to the high syntheticity of scientific texts may be the more conservative nature of this text type. While authors of literary or journalistic texts often have incentives to innovate and introduce novel constructions and expressions to their works, scientists and scholars are less motivated to do so in order to make their product more appealing to the reader. This may be because scientific publications are valued much more for their content rather than for their form. One rarely reads a scientific publication as a form of entertainment or for mere amusement. The reader of such texts is usually interested in facts, arguments, descriptions, and explanations. Therefore, the author of scientific literature is less motivated to introduce novel expressions and constructions. Additionally, academic institutions value traditions and the experience of the past, and this is likely to also influence the writing style employed among academics, allowing for older linguistic expressions to be used longer. German has been said to be losing some of its synthetic forms over time (cf. Rosenberg 2003: 11; Polenz and Wolf 2009: 12), which implies that older expressions contain more inflection. Consequently, the use of older, more conservative expressions in the scientific writing style could be a reason for the higher syntheticity of these texts.

Comparing the results from the corpus of scientific texts from the 21st century to the corresponding corpus for the 20th century, we can see both a higher mean AI (440) and a higher mean SI (494) in the older texts. These results support the aforementioned observations made by Rosenberg (2003) and Polenz and Wolf (2009) about the disappearance of some synthetic constructions, as we see a declining syntheticity over time. At the same time, however, the analyticity of scientific texts also decreases, although to a lesser extent (a 20-index-point decline in SI vs. a 14-index-point decline in AI). Compared to the two corpora with literary texts, the differences between the 20th and 21st centuries here are more substantial, which implies that the scientific writing style has changed more than the literary writing style over the past 100 years. This may be due to expansion and standardization processes that have taken place in academic institutions in the latter half of the 20th and in the first two decades of the 21st century. One should also not rule out the influence of English, a much less inflecting language than German, as a lingua franca of the sci-



entific community and the strong exposure of academics to it through conferences, publications and popular culture.

Similar to what was observed for the corpora containing literary texts, one notices a decline in the mean GI. In this case, however, the decrease is stronger, amounting to 34 index points (from 934 in the 20th-century corpus to 900 in the 21st-century corpus). Looking at the results from the 21st-century corpora, the scientific text type is also the one that encodes the most grammatical information (having the highest GI). This observation appears consistent with the arguments made above about the relative conservativeness and adherence to established tradition, characteristic of the scholarly and scientific writing styles. However, it has to be noted that, in the 20th-century corpora, it is the journalistic texts that exhibit the highest level of grammaticity. The diachronic GI decline in the scientific text type also suggests that some changes have been underway over the past 100 years in the writing style and language of the scholarly and scientific community.

### 3.3. Journalistic Text Type

Moving on to the corpus of 21st-century journalistic texts, we observe results that are similar to those registered in the corpora of scientific texts, yet the disparity between AI and SI here is not that large. This corpus yielded the lowest mean AI (415) of the three corpora of 21st-century texts (see Table 1) and a rather medium mean SI (448). The low AI score may be explained by the absence of dialogue<sup>5</sup>, with the occasional quote in such texts usually being formal utterances by officials, and by the requirements for brevity and linguistic economy that are sometimes imposed on reports and articles to save both printing space and reading time. The SI score of 448 is somewhere between the results from the other two corpora of 21st-century texts (literary corpus – 415, scientific corpus – 474). This may be associated with the fact that journalistic texts are not as focused on narrative and dialogue as prose literary texts are, while they are also not as prone to a more conservative writing style as academic publications are.

Taking a look at the results from the corpora of journalistic texts from the 20th century, we see some substantial differences, especially in the mean SI column (see Table 1 and Table 2). The mean SI for journalistic texts in the 20th-century corpus amounts to 523, which is 75 index points more than the result for the 21st-century journalistic texts. This major difference shows that, with respect to its morphological structure, the journalistic writing style has changed significantly over the past 100 years. The reasons for this change could be many, and given the available data in this study, it can only be speculated what the main driving force is. Yet, one observation made during the analysis of the individual corpus samples may provide

---

<sup>5</sup> Note, though, that the corpora of journalistic texts do not include full-fledged interviews, which are precisely the type of journalistic texts that involve dialogue, although certainly not as informal and loose as the one encountered in literary works.

some direction. Many of the 20th-century newspaper articles that were digitized and analyzed for this study expressed at least some amount of personal political views and offered commentary, attempting to be not only informative, but also persuasive. The texts included in the 21st-century journalistic corpus did not offer as much commentary and personal opinions, but were rather more neutral, focusing on being informative rather than persuasive<sup>6</sup>. Here, too, a detailed part-of-speech analysis of the samples could provide insights into the reasons for the observed decrease in syntheticity in journalistic texts.

The analysis of the two journalistic corpora also yielded some interesting observations about the grammaticity of journalistic texts from the 20th and 21st centuries. While the scientific corpus is the one with the highest mean GI from among the corpora of 21st-century texts, it is the journalistic corpus that holds the first place with a mean GI of 963 among those comprising 20th-century texts. Comparing this result to the corresponding score from the 21st-century corpus, a decline of 100 index points can be measured – the highest difference between any two of the mean indices measured for the corpora analyzed. The suggestions provided above to explain the significant decrease in syntheticity of the journalistic texts could also apply to the changes measured in this metric, especially considering that the syntheticity index is indeed one of its components.

### **3.4. Diachronic Comparison of the Overall Results**

As can be seen in Table 1 and Table 2, the overall mean AI and SI have been calculated for all 21st-century corpora on the one hand, and for all 20th-century corpora on the other hand. This was done in order to allow comparison across the corpora and thus to draw overall conclusions about the development of the morphological-typological profile of the German language as a whole, to the extent that the three investigated text types can be representative for the entire language system. It should be acknowledged, however, that written German includes many other text types apart from the ones investigated here. For this reason, the conclusions offered here should be viewed as suggestions that need to be substantiated through the investigation of more text types, and especially of the spoken language.

A brief look at the scores shows that both the overall mean AI and the overall mean SI of the 20th-century corpora are higher than their 21st-century counterparts. These measurements indicate that, over the course of the 20th and the first two decades of the 21st century, German has simultaneously become less analytical and less synthetic from a morphological point of view. However, it has to be noted that the decrease in syntheticity (35 index points) is more than twice as strong as the decrease in analyticity (14 index points). This result seems to partially align with

---

<sup>6</sup> It should be noted that the 21st-century journalistic corpus does include opinion pieces and political analyses, but none of them ended up in the randomly selected samples that were analyzed.

observations made by a number of linguists ever since the 19th century, who point out that there are tendencies towards greater analyticity and lesser syntheticity in many Indo-European languages, including German (cf. Schlegel 1818: 17; Tristram 2009: 255; Haspelmath and Michaelis 2017: 2). While a declining syntheticity could indicate a process of analytization, however, the decrease in analyticity observed in this analysis certainly does not support such a conclusion. In fact, the results from the current analysis indicate that the concepts of analyticity and syntheticity must not necessarily stand at the two ends of a continuum, and that a shift away from one of them must not necessarily entail a move towards the other.

Another interesting observation regarding all corpora involves the overall mean GI. A brief look at Tables 1 and 2 reveals that there is a difference of 49 index points between the two scores (20th-century corpora – 927; 21st-century corpora – 878). These results suggest that the amount of grammatical information encoded in German written texts has declined over the past 100 years. Of course, to solidify this conclusion, further research and testing will be necessary, as the data used for this study come from a total of only 30 samples. In addition, a similar analysis of the spoken language will also be needed if the conclusion is to be expanded to the entire language system, as speech makes up a large part of what can be considered an individual language. Furthermore, a qualitative investigation of the corpus texts could provide additional evidence for the observed grammaticity decline, while a comparison of the diachronic GI development in German to that of other Indo-European languages could yield interesting results from a general typological perspective and provide potential explanations.

#### **4. Conclusion**

This study aimed to explore some differences of the morphological-typological profile of German across three text types, and also to track how this profile has changed over the past 100 years. The most significant findings include: the consistently high syntheticity of scientific texts, possibly stemming from the more conservative nature of this text type; the high syntheticity in 20th-century journalistic texts, probably caused by what was perceived as a more politically and ideologically loaded content and by the more persuasive nature of the articles from the investigated 20th-century period compared to the modern period; the general diachronic decline in both analyticity and syntheticity for all analyzed texts, which implies that analyticity and syntheticity are not necessarily at the two ends of a single continuum; and the consequent decline in grammaticity, a metric combining the AI and SI measurements, which suggests that the amount of grammatical information that is being encoded by language users within the framework of written German has declined over the 20th and the first two decades of the 21st centuries. These conclusions indicate that, from a quantitative perspective, the morphological-typological profile of the German language is continuing to undergo noticeable changes in the

21st century. The precise causes for this process, as well as the factors that play a role in defining the direction of the changes, remain a subject of speculation for the time being, and further research will be necessary to identify them.

Table 1

Corpus	Sample 1		Sample 2		Sample 3		Sample 4		Sample 5		Mean		Mean GI
	AI	SI	AI	SI	AI	SI	AI	SI	AI	SI	AI	SI	
Literary corpus, 21st century	434	522	432	398	518	388	440	366	454	402	<b>456</b>	<b>415</b>	871
Scientific corpus, 21st century	396	452	445	454	428	505	443	491	417	467	<b>426</b>	<b>474</b>	900
Journalistic corpus, 21st century	401	498	420	439	429	433	426	428	401	440	<b>415</b>	<b>448</b>	863
<b>Overall</b>											<b>432</b>	<b>446</b>	<b>878</b>

Table 2

Corpus	Sample 1		Sample 2		Sample 3		Sample 4		Sample 5		Mean		Mean GI
	AI	SI	AI	SI	AI	SI	AI	SI	AI	SI	AI	SI	
Literary corpus, 20th century	447	464	483	459	428	365	467	451	462	386	<b>457</b>	<b>425</b>	882
Scientific corpus, 20th century	409	424	424	531	479	430	453	546	437	540	<b>440</b>	<b>494</b>	934
Journalistic corpus, 20th century	444	497	407	479	443	561	441	516	463	564	<b>440</b>	<b>523</b>	963
<b>Overall</b>											<b>446</b>	<b>481</b>	<b>927</b>

## BIBLIOGRAPHY

**Anthony 2022:** Anthony, L. *TagAnt* (Version 1.3.0) [Computer Software]. Tokyo, Japan: Waseda University.

**Bowker and Pearson 2002:** Bowker, L., J. Pearson. *Working with Specialized Language: A practical guide to using corpora*. London: Routledge.

**Brinker 2005:** Brinker, K. *Linguistische Textanalyse: eine Einführung in Grundbegriffe und Methoden*. Berlin: Erich Schmidt Verlag.

**Fischer 2009:** Fischer, C. *Texte, Gattungen, Textsorten und ihre Verwendung in Lesebüchern*. Dissertation, Justus Liebig University Giessen.

**Greenberg 1960:** Greenberg, J. A quantitative approach to the morphological typology of language. – In: *International Journal of American Linguistics*, Vol. 26. No. 3, pp. 178 – 194 <[https://archive.org/details/sim\\_international-journal-of-american-linguistics\\_1960-07\\_26\\_3/page/178/mode/2up](https://archive.org/details/sim_international-journal-of-american-linguistics_1960-07_26_3/page/178/mode/2up)> [5.09.2023].

**Haspelmath and Michaelis 2017:** Haspelmath, M. and S. M. Michaelis. Analytic and synthetic: Typological change in varieties of European languages. – In: Buchstaller, I. and B. Siebenhaar (eds.), *Language Variation: European Perspectives VI: Selected Papers from the Eighth International Conference on Language Variation in Europe (ICLaVEe 8)*, Amsterdam, The Netherlands: John Benjamins Publishing Co., pp. 3 – 22.

**Horsch 2021:** Horsch, J. Typological profiling of English, Spanish, German and Slovak: A corpus-based approach. – In: *Journal of Linguistics/Jazykovedný časopis* Vol. 72. No. 2, pp. 342 – 352, [online] doi:10.2478/jazcas-2021-0032 [5.09.2023].

**Kock 1997:** Kock, J. Gramática y corpus: los pronombres demostrativos. – In: *Revista de filología románica*, Vol. 14. No. 1, pp. 291 – 298.

**Marty 1908:** Marty, A. *Untersuchungen zur Grundlegung der Allgemeinen Grammatik und Sprachphilosophie*. Halle: Niemeyer.

**Polenz, Wolf 2009:** Polenz, P., N. R. Wolf. *Geschichte der deutschen Sprache*, 10. Ed., Berlin/New York, Germany/USA: Walter de Gruyter.

**Rosenberg 2003:** Rosenberg, P. Vergleichende Sprachinselforschung: Sprachwandel in deutschen Sprachinseln in Russland und Brasilien. – In: *Linguistik online*, Vol. 13. No. 1, [online] doi:10.13092/lo.13.881 [5.09.2023].

**Schlegel 1818:** Schlegel, A.W. *Observations sur la langue et la littérature provençales*, Paris, France: Librairie grecque-latine-allemande  
<<https://gallica.bnf.fr/ark:/12148/bpt6k63239579/f24.item.texteImage#>> [5.09.2023].

**Schmid 1994:** Schmid, H. *Improvements in Part-of-Speech Tagging with an Application to German. Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland  
<<https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger2.pdf>> [5.09.2023].

**Schmid 1995:** Schmid, H. *Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK < > [5.09.2023].

**Seghiri 2014:** Seghiri, M. Too big or not too big: Establishing the minimum size for a legal ad hoc corpus. – In: *Hermes – Journal of Language and Communication in Business*, Vol. 53 pp. 85 – 98.

**Szmrecsányi 2009:** Szmrecsányi, B. Typological parameters of intralingual variability: Grammatical analyticity versus syntheticity in varieties of English. – In: *Language Variation and Change*, Vol. 21, pp. 319 – 353.

**Szmrecsányi 2016:** Szmrecsányi, B. An analytic-synthetic spiral in the history of English. – In: E. van Gelderen (ed.), *Cyclical Change Continued*, pp. 93 – 112, [online] doi:10.1075/la.227 [5.09.2023].

**Szmrecsányi, Kortmann 2011:** Szmrecsányi B., B. Kortmann. Typological profiling: Learner Englishes versus indigenized L2 varieties of English. – In: J. Mukherjee and M. Hundt (eds.), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a paradigm gap*, Amsterdam/Philadelphia, The Netherlands/USA: John Benjamins Publishing Co., pp. 167 – 188.

**Tristram 2009:** Tristram, H. Wie weit sind die inselkeltischen Sprachen (und das Englische) analytisiert? – In: U. Hinrichs and N. Reiter and S. Tornow (eds.), *Eurolinguistik: Entwicklung und Perspektiven*, Wiesbaden, Germany: Harrassowitz, pp. 255 – 280.