Journal of Siberian Federal University. Humanities & Social Sciences
2020 13(12): 2035–2048

DOI: 10.17516/1997-1370-0702
УДК 81'42

Corpus Analysis of the Grammatical Categories' Constituents in Fiction Texts Considering the Linguo-Regional Component

Alexey I. Gorozhanov and Innara A. Guseynova*
Moscow State Linguistic University
Moscow, Russian Federation

Received 27.03.2020, received in revised form 08.09.2020, accepted 10.12.2020

Abstract. The problem of insufficient development of professional digital tools in the humanities, including linguistics, is posed. The results of a long-term study on the development of specialized software for corpus analysis of a foreign language text, in which the regional component plays an important role, are described. On the example of Franz Kafka's novel "Das Schloss" the work of three computer programs is demonstrated, including primary text processing, automatized markup of a small special (grammatical) corpus, a combined search for selected constituents of the grammatical categories of certainty / uncertainty, modality, as well as individual grammatical phenomena by markers of the marked up corpus and by regular expressions in case of the not marked up text, considering the regional characteristics of spelling. It is concluded that the maximum effectiveness of research of this kind can be achieved with the active work of an expert with a specially developed software product. Moreover, the expert's task includes not only professional software management, but also active participation directly in its development. The described analysis model can be used to analyze the categorical spectrum of a foreign language text, as well as to compile didactic materials for teaching foreign languages and translation.

Keywords: corpus analysis, grammatical categories, special software, linguo-regional component, cross-cultural communication, digitalization, German language, Franz Kafka, regular expressions.

Research area: linguistics.

Citation: Gorozhanov, A.I., Guseynova, I.A. (2020). Corpus analysis of the grammatical categories' constituents in fiction texts considering the linguo-regional component. J. Sib. Fed. Univ. Humanit. Soc. Sci., 13(12), 2035–2048. DOI: 10.17516/1997-1370-0702.

* Corresponding author E-mail address: a.gorozhanov@linguanet.ru, i.guseynova@linguanet.ru
ORCID: 0000-0003-2280-1282 (Gorozhanov); 0000-0002-6544-699X (Guseynova)

Korpusanalyse der Konstituenten Grammatischer Kategorien im Literarischen Text mit Berücksichtigung der Linguoregionalen Komponente

Alexey I. Gorozhanov and Innara A. Guseynova*
Moscow State Linguistic University
Moscow, Russian Federation

Annotation. Im Beitrag wird das Problem der unzureichenden Digitalisierung der Geisteswissenschaften, einschließlich der Linguistik, aufgeworfen. Beschrieben werden die Ergebnisse einer sieben Jahre dauernden Forschung zur Entwicklung einer speziellen Software zur Korpusanalyse eines fremdsprachigen Textes, in der die regionale Komponente eine wichtige Rolle spielt. Am Beispiel des Romans "Das Schloss" von Franz Kafka wird die Arbeit von drei Computerprogrammen gezeigt, darunter die primäre Textverarbeitung, das automatisierte Markup eines kleinen speziellen (grammatischen) Korpus, die kombinierte Suche nach den ausgewählten Konstituenten der grammatischen Kategorien "Bestimmtheit/Unbestimmtheit" und "Modalität" sowie einzelner grammatischer Phänomene durch die Attribute des markierten Korpus und durch reguläre Ausdrücke für den nicht markierten Text mit Berücksichtigung der regionalen Merkmale in der Rechtschreibung. Es wird geschlussfolgert, dass eine maximale Wirksamkeit derartiger Forschung durch die aktive Arbeit eines Experten mit und an diesem speziell entwickelten Softwareprodukt erreicht werden kann. Darüber hinaus umfasst die Aufgabe des Experten nicht nur den professionellen Umgang mit der Software, sondern auch die aktive Teilnahme direkt an seiner Entwicklung. Das entwickelte Analysemodell kann verwendet werden, um das kategoriale Spektrum eines fremdsprachigen Textes zu bestimmen sowie didaktische Materialien für den Fremdsprachenunterricht und das Übersetzungstraining zusammenzustellen.

Schlüsselwörter: Korpusanalyse, grammatische Kategorien, Spezialsoftware, linguoregionale Komponente, interkulturelle Kommunikation, Digitalisierung, deutsche Sprache, Franz Kafka, regulärer Ausdruck.

Einleitung

Die globale Orientierung am Digitalen, d. h. an der Digitalisierung aller Aspekte unseres Lebens erfordert die Entwicklung erstklassiger Spezialsoftware.

Heutzutage werden technologische Lösungen auf dem Gebiet des technischen und naturwissenschaftlichen Wissens bevorzugt, während die Digitalisierung des geisteswissenschaftlichen Wissens abstrakter Natur ist und ohne sichtbare Ergebnisse diskutiert wird, was die Rolle und den Wert von Leistungen auf diesem Gebiet der Wissenschaft herabsetzt.

Trotz der Tatsache, dass die Digitalisierung von Bibliotheksbeständen, die Verwendung von Autoguides oder die Erstellung von Präsentationen weit verbreitet ist, werden Probleme der Linguistik, Literaturwissenschaften, Regionalwissenschaften und anderer Geisteswissenschaften nur unzureichend berücksichtigt. Aus diesem Grund werden die Ressourcen der Informations- und Kommunikationstechnologien nicht voll ausgeschöpft, und Computerlösungen im geisteswissenschaftlichen Bereich sind oft unprofessionell oder sogar primitiv.

Theoretischer Rahmen

Eine Analyse der Fachliteratur zeigt, dass sowohl Arbeiten relativ junger Wissenschaftler, die noch nicht bekannt geworden sind, als auch grundlegende Arbeiten erfahrener Wis-

senschaftler den Fragen der Digitalisierung gewidmet sind (Ryndina, 2019; Bäckermann, 2018; Potapova, Potapov, 2015). In unseren Werken haben wir versucht, die vorhandenen Erfahrungen und Ergebnisse unter Berücksichtigung ihrer praktischen Bedeutung und Durchführbarkeit zu systematisieren, um den Vektor für die Implementierung und Verbreitung von Informations- und Kommunikationstechnologien im geisteswissenschaftlichen Bereich zu bestimmen (Gorozhanov et al., 2018; Gorozhanov, 2019). Im Zusammenhang damit erscheint es uns angebracht, die Entwicklung der theoretischen und angewandten Grundlagenforschung mit ihrer anschließenden Implementierung im virtuellen Raum zu intensivieren.

In Bezug auf Sprachkenntnisse weisen wir auf die Notwendigkeit hin, linguistische theoretische und technische Instrumente zu systematisieren, die für den Einsatz in der Digitalisierung geeignet sind. In diesem Artikel versuchen wir, die von uns im Rahmen der Korpusanalyse entwickelten Sprachinstrumente zu testen. Die Korpusanalyse ist eine Forschungsmethode, die in unserer Zeit nicht an Popularität verloren hat und zur Lösung verschiedener Sprachprobleme verwendet wird. Dazu gehören die Fremdsprachendidaktik und das Übersetzungstraining (Giampieri, 2020; Silva, 2020); die Untersuchung einzelner sprachlicher Phänomene (Duque, 2020; Davies et al., 2020; Meidani, 2019) und selbst das Studium philosophischer Fragen (Caton, 2020).

Problemstellung

Jede Software wird entwickelt, um ein bestimmtes Ziel zu erreichen, das die Vorbereitung und Anwendung einzigartiger Lösungen erfordert. Im Fall des Sprachkorpus – beginnend mit dem allerersten computergestützten Korpus "The Brown University Standard Corpus of Present-Day American English", der 1963 erschien (Mordovin, 2013: 51) – beschäftigen sich Wissenschaftler mit universellen Entwicklungen, mit deren Hilfe eine Vielzahl von theoretischen und angewandten Problemen gelöst werden soll. Die Linguistik und die Sprachwissenschaft im Allgemeinen arbeiten mit Textkorpora, deren Verarbeitung in der Analyse unter Berücksichtigung der Spezifik des Sprachmaterials besteht. Mit anderen Worten, Computerlösungen sind für moderne Sprachforschungen unabdinglich, einschließlich zur Bestimmung regionaler Merkmale (zum Beispiel in Dialektforschungen). Die Korpuslinguistik verfügt über alle notwendigen Instrumente und "zeichnet sich im Allgemeinen im Paradigma der modernen Sprachforschung durch methodische Universalität und Wirksamkeit aus" (Barkovich, 2016: 5).

Diese Behauptung scheint zweifellos richtig zu sein, aber ein Sprachkorpus, das für die Sprachforscher von unschätzbarem Wert ist, ist nicht immer wirksam, um ein hochspezialisiertes Problem zu lösen. In unserem Fall verfolgen wir das Ziel, ein sowohl für einen markierten als auch für einen nicht markierten Text geeignetes Instrument zur automatisierten Analyse des Textes eines literarischen Werkes in einer Fremdsprache zu schaffen, um die Mittel zum Ausdruck grammatischer Kategorien zu bestimmen. Es wird davon ausgegangen, dass flexiblere Suchanfragen durch die Arbeit an einem nicht markierten Text gestellt werden können, die beispielsweise regionale Merkmale berücksichtigen. Eine automatisierte Analyse eines fremdsprachigen Textes ist hilfreich bei der Zusammenstellung von didaktischen Materialien auf der Grundlage dieses Textes mit dem Ziel, den Studierenden die Sprache als Ganzes sowie deren Aspekte, einschließlich territoriale Besonderheiten, zu vermitteln.

Technologisch gesehen ist die Software ein Paket von Computerprogrammen, die in einer modernen, vielversprechenden und auf nichtkommerzieller Basis verteilten Programmiersprache geschrieben werden, um eine weitere Modernisierung der Pilotversion und deren Tests vor einer umfassenden Implementierung zu ermöglichen. Die im Paket enthaltenen Programme funktionieren offline und sind für die Hardware nicht anspruchsvoll.

Methodologie

Die Verwendung von Standardsoftware hat sowohl Vor- als auch Nachteile. Letztere bestehen vor allem darin, dass sich eine solche Software auf die Arbeit mit einem abstrakten Text konzentriert, der keine eigene Origina-

lität und keinen eigenen künstlerischen Wert hat. Eine solche Software ist zweifellos für die Arbeit mit technischen Texten geeignet und orientiert sich am Suchen und Erkennen von Termini, führt jedoch keine analytische Verarbeitung des Textes durch, die der Interpretation vorausgeht. Die Analyse linguoregionaler Komponenten ist noch schwieriger, da das Sprachkorpus in der Regel auf der Grundlage einer Literatursprache, aber nicht auf der Grundlage von Dialekten erstellt wird. Darüber hinaus erfordert eine solche Software eine ständige Anpassung an die tatsächlichen Bedürfnisse der Benutzer. Eine teilweise Verbesserung ist möglich, löst jedoch nicht das Problem einer umfassenden Textanalyse. Unserer Meinung nach müssen die entwickelten Programme verschiedene Aufgaben einer Korpusanalyse erfüllen – von der anfänglichen Textverarbeitung und seiner Markierung mit dem Ziel, ihn in ein Sprachkorpus umzuwandeln, bis zur mehrstufigen Suche im erstellten Korpus und Speichern der Ergebnisse in einem benutzerfreundlichen Format. Darüber hinaus soll die Suche auch mit einem nicht markierten Text funktionieren.

Aus diesem Grund soll das Softwarepaket mindestens drei Komponenten enthalten:

– ein Programm zur primären Textverarbeitung;

– ein Programm zur Umwandlung dieses Textes in ein Sprachkorpus (Korpuseditor);

– einen Korpusmanager für die Arbeit mit einer markierten (und nicht markierten) Variante.

Es wird vorausgesetzt, dass der sogenannte "rohe" Text des literarischen Werkes eine Textdatei (TXT-Format) ist. Als Ergebnis der Softwareverarbeitung soll daraus eine Datenbank werden (z. B. im XML-Format). Die zweite Komponente des Softwarepakets, d. h. der Korpuseditor, arbeitet mit dieser Datenbank. Es muss betont werden, dass das XML-Format praktisch ist, weil es universell und allgemein anerkannt ist. Darüber hinaus ist es relativ einfach und kann nicht nur programmgesteuert, sondern bei Bedarf auch "manuell", mit einem beliebigen Texteditor, gelesen werden. Jede Datenbankzeile im XML-Format erhält eine unikale Nummer, beginnend mit "eins".

In unserem Fall ist die grundlegende Struktureinheit des Korpus ein Satz (nicht das Wort), wodurch die Marker der syntaktischen Ebene berücksichtigt werden können. Die grammatischen Marker im Satz unterliegen ebenfalls einem Markup. Daher bestimmen wir unser Sprachkorpus als "klein" und "speziell" (grammatisch).

Die XML-Datei muss folgende Attribute enthalten:

– Charakteristik des Satzes: einfacher Satz, Satzgefüge oder Satzreihe;

– Zeitform des Prädikats;

– Modus des Prädikats: Indikativ, Imperativ oder Konjunktiv;

– Vorhandensein eines Modalverbs;

– Vorhandensein eines Artikels: Nullartikel, bestimmter Artikel oder unbestimmter Artikel;

– Vorhandensein von Präpositionen: mit dem Genitiv, dem Dativ, dem Akkusativ oder der doppelten Rektion;

– Vorhandensein eines Verbs mit einem trennbaren oder untrennbaren Präfix;

– Vorhandensein von Adjektiven im Positiv, im Komparativ oder im Superlativ;

– Vorhandensein eines Prädikativs (eines zusammengesetzten Prädikats);

- Genus des Prädikats: Aktiv, Passiv oder Stativ;

- Vorhandensein reflexiver Verben.

Die Wahl dieser Attribute beruht auf der Tatsache, dass das Sprachkorpus als Instrument für die Erarbeitung didaktischer Materialien verwendet werden soll. Unter anderem wurden diejenigen grammatischen Phänomene ausgewählt, die Schwierigkeiten beim Erlernen der deutschen Sprache bereiten, insbesondere beim Übersetzen aus dem Deutschen ins Russische. Die ausgewählten Attribute sind leicht zu markieren, da das Markieren des Textes automatisiert erfolgt, d. h. der Benutzer entscheidet über das Vorhandensein oder Fehlen des einen oder anderen Markers.

Der Korpuseditor bedarf eines grafischen Benutzerinterface. Der folgende Algorithmus zum Markieren des Korpus wird angenommen. Über das grafische Menü wird die XML-Datenbankdatei geladen. Der erste Satz, d. h. die erste Reihe in der Datenbank, wird in einem

speziellen Fenster angezeigt. Mithilfe von Widgets werden die Attribute markiert, die im laufenden Satz vorhanden sind. Nach allen Markierungen geht das Programm durch das Drücken des Knopfes "Weiter" zum nächsten Satz über. Nach Abschluss der Arbeit wird das Resultat über das Grafikmenü in der Datenbankdatei gespeichert. Es ist auch möglich, die Datenbank unter einem anderen Namen zu speichern. Zusätzlich zum Knopf "Weiter" soll es auch den Knopf "Zurück" geben. Für den Übergang zu irgendeinem Satz der aktuellen Datenbank muss im Menü die Funktion "GoTo [Satznummer]" aktiviert werden.

Ganz unten im Interface soll sich eine Statusleiste befinden, die den Benutzer über die von ihm letzte ausgeführte Aktion informiert, z. B. das Speichern einer Datei oder das Öffnen einer neuen Datenbank.

Nachdem der Text vollständig markiert und die Datenbank gespeichert worden ist, muss ein Verfahren für die Arbeit damit bereitgestellt werden. Diesem Zweck soll das dritte Programm des Softwarepakets, nämlich das Korpussuchsystem oder der Korpusmanager, dienen.

Dies ist das komplexeste Programm in seiner Struktur und Funktionalität. Es wird davon ausgegangen, dass es in zwei Modi funktioniert. Im ersten Modus wird mit einem markierten Sprachkorpus unter Verwendung markierter Attribute gearbeitet. Im zweiten Modus wird mit einem nicht markierten Text gearbeitet, der jedoch mit dem ersten Programm des Softwarepakets erstellt wurde und auch eine XML-Datenbank darstellt. Somit kann man im zweiten Modus schon vor dem vollständigen Markup (durch das zweite Programm des Softwarepakets) mit der Datenbank arbeiten. Wenn die Datenbank bereits markiert ist, soll der zweite Modus weiterhin voll funktionsfähig sein.

Das Interface des Korpusmanagers soll ein funktionierendes Output-Fenster (in dem das Suchergebnis platziert wird), sowie ein Block für die Einstellung der Attribute und die Widgets zur Auswahl des zu untersuchenden Textes haben. Im zweiten Modus soll der Block für die Einstellung der Attribute durch Suchsteuerelemente ersetzt werden.

Um das Programminterface ergonomischer zu gestalten, sollen sich die Widgets des ersten und des zweiten Modus auf verschiedenen Tabs befinden. Der dritte Tab kann ein Protokoll enthalten, in dem alle vom Benutzer im Rahmen dieses Programmstarts durchgeführten Aktionen automatisch protokolliert werden sollen.

Da die Verarbeitung von Textdaten nach vielen Parametern lange dauern kann, soll das Programm über einen Fortschrittsbalken verfügen, der den Arbeitsaufwand in Prozent angibt.

Wie im zweiten Programm des Softwarepakets soll sich hier am unteren Rand des Interface eine Statusleiste befinden, die den Benutzer über die zuletzt ergriffenen Maßnahmen informiert.

Es muss eine Kontinuität zwischen den beiden Modi beachtet werden. Dies bedeutet, dass der Benutzer nach dem Starten der Suche im ersten Modus in den zweiten Modus wechseln und die Arbeit mit den Ergebnissen fortsetzen kann, die während der Verwendung des ersten Modus erzielt wurden. Angenommen, im ersten Modus wurden alle Satzgefüge gefunden. Wenn der Benutzer in den zweiten Modus wechselt, kann er alle Sätze finden, die die eine oder andere untergeordnete Konjunktion enthalten. Die Sätze, die die Suchkriterien nicht erfüllen, sollen auf Wunsch des Benutzers mit einem speziellen Knopf aus dem Output-Fenster ausgeschlossen werden. Als Ergebnis der Suche werden somit alle Satzgefüge eines bestimmten Typs gefunden.

Der zweite Modus soll die mehrstufige Suche unterstützen. Wenn wir das obige Beispiel fortsetzen, kann der Benutzer eine bestimmte lexikalische Einheit oder eine Reihe von lexikalischen Einheiten in den Satzgefügen eines bestimmten Typs finden.

Suchergebnisse im zweiten Modus sollen mit einer Hintergrundfarbe gekennzeichnet sein, und es soll die Möglichkeit bestehen, diese Farben zu wechseln. Bei der Suche nach untergeordneten Konjunktionen wird beispielsweise eine bestimmte Farbe ausgewählt, bei der Suche nach lexikalischen Einheiten – eine andere. Die Suchmuster des zweiten Modus sollen zur schnellen Wiederverwendung in einer speziellen Datei gespeichert werden.

Eine weitere wichtige Forderung an das dritte Programm des Softwarepakets ist die Möglichkeit, das Ergebnis aus dem Suchfenster in einer Datei in einem universellen Format, beispielsweise HTML, zu speichern.

Diskussion

Geleitet von den oben genannten technischen Spezifikationen haben wir im Zeitraum von 2014 bis 2020 drei Softwareprodukte entwickelt und getestet, die zusammen das erforderliche Softwarepaket bilden.

Als Instrument der Entwicklung wurden die Programmiersprache Python3 und die PyQt5-Grafikbibliothek gewählt, ein universelles Softwareentwicklungstool mit einem grafischen Benutzerinterface, das auf nahezu jedem Betriebssystem funktioniert.

Das erste Programm des Softwarepakets verfügt über kein grafisches Benutzerinterface, weil es nur eine Aufgabe erfüllt – das Konvertieren einer Textdatei in eine XML-Datenbank. Bei der Eingabe erhält das Programm eine Datei, in der der Text in nummerierte Sätze unterteilt ist. Die erste Zeile der Quelldatei ist nicht nummeriert und enthält den Namen des Autors und den Titel des Werkes.

Beim Generieren einer XML-Datei weist das Programm jedem Satz Standardattributwerte zu, die die häufigsten Phänomene der Sprache kennzeichnen (z. B. Aktiv und Indikativ als die häufigsten Genus Verbi und Modus). Dieser Schritt soll in Zukunft die "manuelle" Markierung des Korpus erleichtern.

Das zweite Programm des Softwarepakets – der Korpuseditor – erhielt ein grafisches Benutzerinterface, das aus folgenden Komponenten (von oben nach unten) besteht:

– Menüblock;
– Zeile des Namens der geladenen Datenbank;
– Zeile der Nummer des laufenden Satzes der geladenen Datenbank;
– Output-Fenster für die Ausgabe von Einheiten des geladenen Korpus;
– Attributenblöcke;
– Knöpfe "<---" und "--->";
– Statusleiste (siehe Abb. 1):

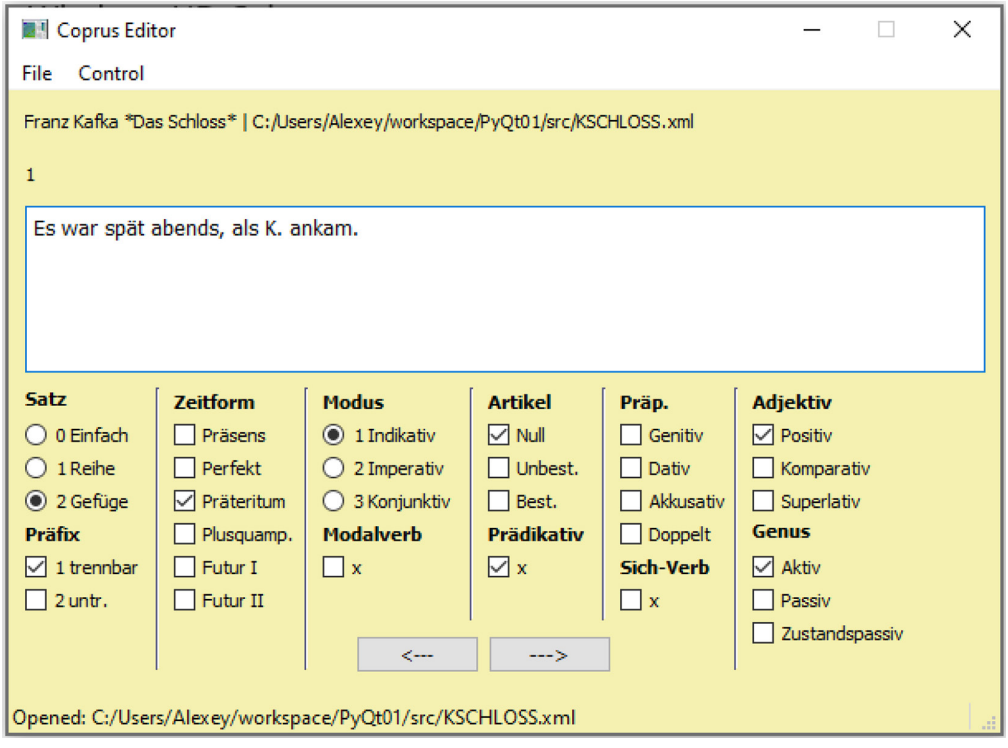

Abb. 1. Interface des Programms mit geladener Datenbank

Der Menüblock enthält zwei Gruppen: "File" und "Control". Die Gruppe "File" enthält vier Punkte: "Open" (öffnet die Datenbankdatei), "Save" (speichert die Datenbankdatei unter dem aktuellen Namen), "Save As" (speichert die Datenbankdatei unter einem neuen Namen) und "Exit" (schließt das Programm). Die Gruppe "Control" hat nur einen Punkt – "GoTo", um zu einem der Sätze der geladenen Datenbank zu gelangen.

Die Zeile für den Namen der Datenbankdatei gibt an, welche Datei derzeit geöffnet ist. Neben dem Autor und dem Titel des Werkes wird der vollständige Name der Datei mit ihrer Adresse angezeigt.

Das Output-Fenster zur Anzeige von Einheiten des geladenen Korpus enthält den aktuellen Datenbankeintrag (Satz). Die grammatischen Markierungen befinden sich in elf Blöcken. Jeder Block hat einen Titel und eine Reihe von Widgets – Kontrollkästchen oder Optionsfeldern.

Der Block "Satz" enthält drei Optionsfelder, sodass der Benutzer nur ein Merkmal des Satzes auswählen muss (einfacher Satz, Satzreihe oder Satzgefüge). Es wird angenommen, dass der Benutzer die eine oder andere Entscheidung trifft, wenn mehrere Merkmale gleichzeitig vorhanden sind (z. B. ein Satz stellt sich als Satzreihe heraus, aber einer seiner Teile ist ein Satzgefüge), je nachdem, welcher Marker dominiert.

Der Block "Präfix" hat zwei Kontrollkästchen, daher gibt es vier mögliche Markierungsoptionen: a) der Satz enthält keine trennbaren oder untrennbaren Verben; b) der Satz enthält mindestens ein Verb mit einem trennbaren Präfix, jedoch kein einziges Verb mit einem untrennbaren Präfix; c) der Satz enthält mindestens ein Verb mit einem untrennbaren Präfix, aber kein einziges Verb mit einem trennbaren Präfix; d) der Satz enthält gleichzeitig ein oder mehrere Verben mit trennbaren Präfixen und ein oder mehrere Verben mit untrennbaren Präfixen.

Der Block "Zeitformen" ist der umfangreichste und enthält sechs Kontrollkästchen entsprechend der Anzahl der Tempora des deutschen Verbs. Hierbei wird eine solche Option berücksichtigt, dass der Satz möglicherweise eine Ellipse sein könnte und überhaupt keine Verbformen enthält. Daher ist das Fehlen von Markierungen in allen sechs Kontrollkästchen zulässig. Andererseits kann ein Satz ein Verb in verschiedenen Zeitformen enthalten, was auch zulässig ist.

Der Block "Modus" ist für die Markierung des Modus verantwortlich und wird durch drei Optionsfelder dargestellt. Zum Beispiel, wenn in einem Satz Verben im Indikativ oder im Konjunktiv gefunden werden, trifft der Benutzer eine Entscheidung auf der Grundlage seiner Vision. Es wird jedoch empfohlen, die Phänomene zu markieren, die am komplexesten und seltensten sind, d. h. wenn es sowohl einen Konjunktiv als auch einen Indikativ gibt, wird "Konjunktiv" als Marker gesetzt.

Der Block "Modalverb" markiert nur das Vorhandensein oder Fehlen von Modalverben im Satz und wird durch ein Kontrollkästchen dargestellt. Eine so einfache Lösung hat einen didaktischen Grund, weil die Modalverben für die Studierenden eine erhebliche Schwierigkeit darstellen.

Der Block "Artikel" markiert mit Hilfe von drei Kontrollkästchen das Vorhandensein oder Fehlen eines Nullartikels, des unbestimmten oder bestimmten Artikels im Satz. Möglicherweise kann ein Satz gefunden werden, der keine Substantive enthält und daher keine Artikel enthalten kann.

Der Block "Prädikativ" hat, wie im Fall von Modalverben, nur ein Kontrollkästchen, das das Vorhandensein oder Fehlen eines zusammengesetzten Prädikats im laufenden Satz anzeigt. Die Notwendigkeit, diese Funktion aufzunehmen, wird auch durch die Fremdsprachendidaktik bestimmt.

Der Block "Präposition" verfügt über vier Kontrollkästchen zum Markieren der Verwendung von Präpositionen mit dem Genitiv, Dativ, Akkusativ oder mit einer doppelten Rektion. Die Kontrollkästchen ermöglichen verschiedene Kombinationen, einschließlich das Fehlen von Markern.

Der Block "Sich-Verb" enthält ein Kontrollkästchen, das das Vorhandensein oder Fehlen eines reflexiven Verbs im Satz anzeigt, was den Studierenden auch Schwierigkeiten bereitet.

Im Block "Adjektiv" mit drei Kontrollkästchen kann man Adjektive im Positiv, Komparativ und Superlativ markieren. Es wird auch die Option berücksichtigt, dass der Satz kein einziges Adjektiv enthält.

Schließlich verfügt der Block "Genus" über drei Kontrollkästchen zum Markieren des Aktivs, Passivs oder Zustandspassivs.

Mit den Knöpfen "<---" und "--->" kann man einen Satz nach links oder rechts im Korpus bewegen. Dies ist praktisch für die sequenzielle Markierung von Sätzen und macht es überflüssig, ständig auf den Menüpunkt "GoTo" zurückzugreifen.

Das dritte Programm des Softwarepakets – der Korpusmanager – erhielt ebenfalls ein grafisches Benutzerinterface, das aus folgenden Komponenten (von oben nach unten und von links nach rechts) besteht:

– Menüblock;
– Suchknopf;
– drei Tabs;
– Behälter für den Inhalt der Tabs;
– Dropdown-Menü zur Auswahl der Datenbankdatei;
– Statusleiste;
– Output-Fenster (siehe Abb. 2):

Der Menüblock enthält zwei Gruppen: "File" und "Settings". Die Gruppe "File" enthält zwei Punkte: "Save As HTML" (speichert den Inhalt des Output-Fensters in einer HTML-Datei, wobei die Formatierung behalten wird) und "Exit". Die Gruppe "Settings" enthält nur einen Punkt – "Color" (um die aktuelle Farbe zum Markieren der Suchergebnisse im zweiten Modus auszuwählen).

Der Knopf "Search" aktiviert die Suche gemäß den eingestellten Parametern in der über das Dropdown-Menü ausgewählten XML-Datei. Die genannten Parameter werden in den ersten beiden Tabs ausgewählt. Betrachten wir den ersten Tab.

Der Tab "Tab 1" sucht in einem markierten Korpus. Die Parameter entsprechen den Korpuseditorblöcken: "Satz", "Zeitform", "Modus", "Präfix", "Präp.", "Modalverb", "Adjektiv", "Artikel", "Prädikativ", "Sich-Verb", "Genus". Jeder der aufgelisteten Blocktitel wird in Form eines Knopfes erstellt, der den entsprechenden Block aktiviert / deaktiviert. An der Suche sind nur aktive Blöcke beteiligt. Weiter gibt es einige Beispiele für die Funktionsweise der Suche:

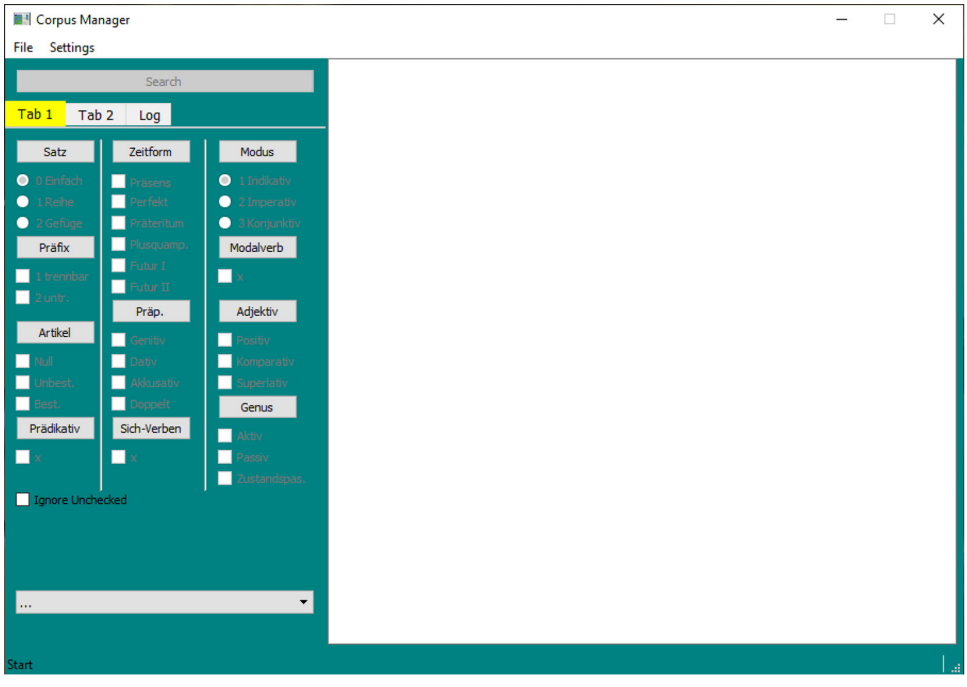

Abb. 2. Interface des Korpusmanagers nach dem Starten des Programms

1. Einstellung: Keiner der Blöcke ist aktiviert. Ergebnis: Die Suche zeigt alle Sätze des Korpus an.

2. Einstellung: Der Block "Satz" ist aktiviert, in dem die Optionsfelder auf "Reihe" gesetzt sind. Ergebnis: Die Suche zeigt Einheiten an, in denen Satzreihen als syntaktischer Hauptmarker festgelegt sind.

3. Einstellung: Der Block "Zeitformen" ist aktiviert, in dem nur das Kontrollkästchen "Perfekt" gesetzt ist. Ergebnis: Die Suche zeigt die Sätze an, die "Perfekt" enthalten (aber keine anderen Zeitformen aufweisen), weil alle anderen Kontrollkästchen dieses Blocks inaktiv sind.

4. Einstellung: Der Block "Zeitformen" ist aktiviert, in dem nur das Kontrollkästchen „Perfekt" gesetzt ist, aber das Kontrollkästchen "Ignore Unchecked" aktiv ist. Ergebnis: Die Suche zeigt die Sätze an, die "Perfekt" enthalten (aber möglicherweise auch andere Zeitformen aufweisen), weil alle inaktiven Kontrollkästchen dieses Blocks ignoriert werden und nicht als erforderlicher Suchparameter betrachtet werden.

5. Einstellung: Aktivierte Blöcke "Modalverb" (mit dem aktiven Kontrollkästchen) und "Satz" mit dem aktiven Optionsfeld "Gefüge". Das Kontrollkästchen "Ignore Unchecked" ist nicht aktiv. Ergebnis: Die Suche gibt Einheiten zurück, die gleichzeitig Modalverben und Satzgefüge enthalten.

In der Statusleiste werden die quantitativen Indikatoren der Suche angezeigt ("N Sätze gefunden").

Durch die Aktivierung verschiedener Blöcke sowie die Einstellungen innerhalb der Blöcke kann man komplexe Suchparameter einstellen. Es muss betont werden, dass im ersten Modus eine maximale Effizienz des Korpus nur erreicht werden kann – wenn es vollständig markiert ist. Andernfalls verfügt jeder Satz über Standardparameter, weil alle Sätze a priori als einfache Sätze mit einem Prädikat im Indikativ angesehen werden (denn in der Gruppe der Optionsfelder muss eine Option immer aktiv sein). Andere Parameter werden nicht markiert.

Der zweite Modus zielt darauf ab, mit einem nicht markierten Korpus zu arbeiten. Der Tab "Tab 2" enthält:

– Eingabefeld für den regulären Ausdruck;

– Startknopf zur Suche nach diesem regulären Ausdruck;

– Knopf zum Ausschließen aller nicht markierten Sätze von den Ergebnissen;

– Dropdown-Menü zur Auswahl eines regulären Ausdrucks aus den Varianten, die in einer TXT-Datei gespeichert sind;

– Knopf zum Hinzufügen eines regulären Ausdrucks aus dem Eingabefeld des regulären Ausdrucks zur Liste in einer TXT-Datei;

– Knopf zum Reinigen von Farbmarkierungen im Output-Fenster (siehe Abb. 3):

Reguläre Ausdrücke werden gemäß den Regeln der PyQt5-Grafikbibliothek erstellt.

Bevor es mit Tab 2 gearbeitet werden kann, müssen in Tab 1 alle Sätze des Korpus im Output-Fenster angezeigt werden, da die Instrumente des zweiten Modus nicht direkt mit dem Korpus arbeiten, sondern mit den Informationen aus dem Output-Fenster. Eine Suche nach regulären Ausdrücken ist praktisch, weil eine große Anzahl von Wortformen in einer einzelnen Abfrage enthalten sein kann. Um beispielsweise alle möglichen Wortformen des Verbs "lassen" zu finden, muss die Abfrage wie folgt eingegeben werden:

```
(\b+([Gg]e)?([Ll][aä]ss)(e|en|t)?\b+)|
(\b+(ließ)(e|en|t|est|et)?\b+)
```

Um nach allen möglichen Formen des deutschen unbestimmten Artikels zu suchen, muss man Folgendes eingeben:

```
\b+[Ee]in(e?)[rsnm]?\b+
```

Textteile, die mit regulären Ausdrücken übereinstimmen, werden mit der aktuellen Hintergrundfarbe markiert (standardmäßig gelb). Im nächsten Suchschritt wird empfohlen, eine andere Farbe zu wählen, um die Ergebnisse des ersten, zweiten Vorgangs usw. zu unterscheiden. Um alle Sätze des Korpus, die keine Farbmarkierung erhalten haben, aus dem Output-Fenster zu entfernen, muss auf den Knopf "Reduce" geklickt werden. Das Endergebnis wird in einer neuen HTML-Datei gespeichert.

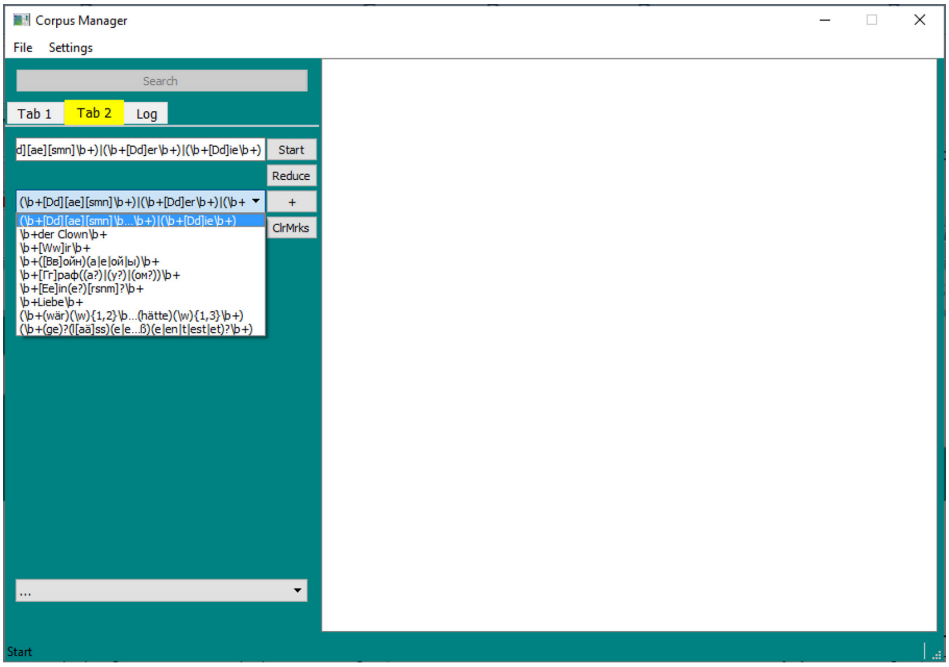

Abb. 3. Tab 2 ist aktiviert

Ein kombiniertes Suchmodell ist effektiv: zuerst im ersten Modus, d. h. über grammatische Parameter und dann über reguläre Ausdrücke im Gefundenen. Die Relevanz der quantitativen und qualitativen Analyse wird beispielsweise von A. Yu. Khomenko hervorgehoben, die feststellt, dass "die qualitative Analyse ermöglicht es, ein vollständiges Modell des Stils des einzelnen Autors (so voll wie möglich) zu erstellen. Und die quantitative Analyse macht dieses Modell objektiver." (Khomenko, 2019: 184).

Die kombinierte Suche ist optimal für die Analyse der Mittel zum Ausdruck der grammatischen Kategorien in einem ausgewählten literarischen Text. Um beispielsweise die zeitliche Struktur mit Hilfe von grammatischen Merkmalen zu analysieren, werden die notwendigen Zeitformen des Verbs hervorgehoben, und mit den Instrumenten von Tab 2 werden die lexikalischen Mittel zum Ausdruck der Temporalität ("heute", "morgen", "letzte" usw.) über reguläre Ausdrücke gesucht.

Weiter folgt ein Beispiel für eine kombinierte Suche. Angenommen, man möchte die Verwendung des Artikels im deutschsprachigen Text fixieren. Wird mit regulären Ausdrücken im zweiten Modus gearbeitet, so ist es einfach, alle Fälle der Verwendung eines bestimmten und unbestimmten Artikels zu finden. Es ist jedoch fast unmöglich, den Nullartikel auf diese Weise zu finden. Im Fall des markierten Sprachkorpus werden alle Verwendungen des unbestimmten Artikels durch einen Klick aufgelistet. Hier könnte man denken, dass der zweite Modus redundant ist, weil auf Tab 1 Optionen für die Verwendung sowohl der Null- als auch der bestimmten und unbestimmten Artikel angezeigt werden könnten. Wenn wir jedoch nicht nur über den Artikel sprechen, sondern allgemein über die Mittel zum Ausdruck der grammatischen Kategorie "Bestimmtheit / Unbestimmtheit", dann fixiert eine Suche nach regulären Ausdrücken beispielsweise Demonstrativpronomen (grenzt sie von den Formen des Artikels ab). Auf diese Weise kann der Prozentsatz der Verwendung des Artikels und der Demonstrativpronomen berechnet werden.

Erfahrungen mit nicht markiertem Textmaterial sind unter anderem aus folgendem Grund wichtig. Das Markieren des Sprach-

korpus, auch wenn es wie in unserem Fall ein kleines spezielles ist, ist äußerst zeitaufwändig. Daher soll man es machen, wenn man sicher ist, dass die Arbeit mit der markierten Variante einen wesentlichen Beitrag zum Ergebnis der Forschung leistet. Vorläufige Informationen, die bei der Entscheidung über die Angemessenheit des Markierens eines Textes helfen, bietet eine Suche im zweiten Modus. Zum Beispiel kann sie die Konzentration des einen oder anderen Mittels zum Ausdruck der Negation in einem literarischen Werk zeigen, basierend auf den Dominanten der Kategorie: "nicht" und "kein".

In diesem Zusammenhang werden wir typische Suchmodelle im zweiten Modus für die vorläufige Bewertung der Mittel zum Ausdruck einiger grammatischer Kategorien am Beispiel des Romans von F. Kafka "Das Schloss" (dargestellt in Form von 4466 nicht markierten Sätzen) betrachten, nämlich der Kategorien "Bestimmtheit / Unbestimmtheit" und der Modalität.

Wir bezeichnen die folgenden Mittel der grammatischen Kategorie der Bestimmtheit / Unbestimmtheit als fixierungsbedürftig in unserem Experiment: Artikel und Pronomen "dieser", "jener", "einige" ("einiger", "einiges"), "jemand", "etwas".

Die Suche erfolgt in mehreren Phasen. In der ersten Phase werden die Sätze fixiert, die Formen eines bestimmten und dann eines unbestimmten Artikels enthalten. Suchergebnis: 7660 Verwendungen des bestimmten Artikels und 2000 Verwendungen des unbestimmten Artikels. Als nächstes wird nach den ausgewählten Pronomen gesucht. Suchergebnisse für "dieser" und "jener" (einschließlich Wortformen): 642 Verwendungen; für "einige" (einschließlich Wortformen): 57 Verwendungen; für "jemand" (einschließlich Wortformen): 59 Verwendungen; für "etwas": 191 Verwendungen. Durch Drücken des Knopfes "Reduce" werden die Sätze im Output-Fenster auf 3525 reduziert. In (4466–3525=) 941 Sätzen vom Korpus des Romans werden diese Konstituenten also überhaupt nicht verwendet. Vermutlich dominiert unter den Artikelwörtern der unbestimmte Artikel. Das erhaltene Ergebnis ermöglicht es uns, zum einen, die ausgewählten Konstituenten nach Häufigkeit zu ordnen und zum anderen, das prozentuale Verhältnis ihrer Verwendung zueinander und zum Text des gesamten Romans mit hoher Genauigkeit zu bestimmen.

Die Protokolldatei für die Suche hat folgenden Inhalt:

1. Read to DOM: KSCHLOSS.xml
2. 4466 item(s) found
3. 7660 match(es) found for pattern: (\b+[Dd][ae][smn]\b+)|(\b+[Dd]er\b+)|(\b+[Dd]ie\b+)
4. 2000 match(es) found for pattern: \b+[Ee]in(e?)[rsnm]?\b+
5. 642 match(es) found for pattern: (\b+[Dd]ies(e[smnr]?)?\b+)|(\b+[Jj]ene[srnm]?\b+)
6. 57 match(es) found for pattern: (\b+[Ee]inig(e[smnr]?)?\b+)
7. 59 match(es) found for pattern: (\b+[Jj]emand(e[mn])?\b+)
8. 191 match(es) found for pattern: (\b+[Ee]twas\b+)
9. Reduced to 3525 item(s)

Die Kategorie der Modalität ist ein schwieriges Objekt für unsere Forschung, weil ihre Konstituenten vielfältig und schwer zu formalisieren sind. In diesem Zusammenhang konzentrieren wir uns auf die Modalverben "müssen", "sollen" und "lassen". Es ist zu beachten, dass wir in diesem Fall absichtlich einen "Fehler" begehen, weil wir auf diese Weise nach Wortformen von Modalverben suchen, unabhängig von ihrer Bedeutung, d. h. indem wir nach regulären Ausdrücken suchen, können wir die modalen Bedeutungen des Verbs "lassen" nicht von seinen nichtmodalen Bedeutungen trennen. Das Modalverb "müssen" kann in bestimmten Fällen durch eine Konstruktion mit dem Verb "nutzen" ersetzt werden, eine Suche ausschließlich nach dem Paradigma des Modalverbs "muss" wird dies jedoch nicht offenbaren (Schäfer, 2004: 133). Das von uns entwickelte Softwarepaket sieht jedoch keine solche Differenzierung im Suchmodus für Tab 1 vor. Dies liegt unter anderem daran, dass die Modalverben für die Studierenden nicht leicht zu verstehen sind. Daher muss jeder Fall der Verwendung des Modalverbs in der Lehrveranstaltung sorg-

fältig analysiert und mit den Studierenden besprochen werden.

Die Suche erfolgt also auch in mehreren Phasen. In der ersten Phase werden alle Verwendungen des Modalverbs "müssen" ermittelt (es ist zu beachten, dass der untersuchte Roman der alten Rechtschreibung folgt und reguläre Ausdrücke die Groß- und Kleinschreibung unterscheiden): "müssen, muß, mußt, müßt; mußte, mußtest, mußten, mußtet; gemußt; müßte, müßtest, müßtet, müßten" und "Müssen, Muß, Mußt, Müßt; Mußte, Mußtest, Mußten, Mußtet; Gemußt; Müßte, Müßtest, Müßtet, Müßten".

Beim Erstellen eines regulären Ausdrucks wird die folgende Methode zur Überprüfung seiner Richtigkeit verwendet. Im Output-Fenster werden "manuell" alle Wortformen platziert, die mit dem aktivierten regulären Ausdruck gefunden werden müssen. Wenn die Suche alle diese Wortformen markiert, wird der reguläre Ausdruck korrekt erstellt.

Die Suchergebnisse für das Modalverb "müssen" (einschließlich Wortformen in der alten Rechtschreibung): 336 Verwendungen. In der zweiten und dritten Phase wird nach den Modalverben "sollen" und "lassen" gesucht (in der alten Rechtschreibung).

Suchergebnisse für das Modalverb "sollen" (einschließlich Wortformen): 201 Verwendungen. Suchergebnis für das Modalverb "lassen" (einschließlich Wortformen in der alten Rechtschreibung): 188 Verwendungen. So sieht der Inhalt der Protokolldatei für die Suche aus:

Start

Read to DOM: KSCHLOSS.xml

4466 item(s) found.

336 match(es) found for pattern: (\b+([Mm]üssen)|([Mm]u(ß|ss)(t|te|test|ten)?)|([Mm]ü(ß|ss)(t|test|en|tet?))|([Gg]emu(ß|ss)t)\b+)

201 match(es) found for pattern: (\b+([Ss]oll(t|st|en|te|test|ten|est|et|e)?)|([Gg]esollt)\b+)

188 match(es) found for pattern: (\b+([Gg]e)?([Ll][aä](ss|ß))(e|en|t)?\b+)|(\b+(ließ)(e|en|t|est|et)?\b+)

Es ist zu beachten, dass die erstellten regulären Ausdrücke sowohl für die alte als auch für die neue Rechtschreibung gültig sind. Wird "ß" durch "ss" ersetzt, so wird zusätzlich die Schreibweise der schweizerischen nationalen Variante der deutschen Sprache, also die regionale Komponente, berücksichtigt.

Weiter folgt ein Beispiel für das Fixieren eines bestimmten grammatikalischen Phänomens: des verbalen Präfixes "über". Die Analyse des Phänomens zeigt, dass die Suche mit regulären Ausdrücken der trennbaren Variante des Präfixes nicht effektiv ist. In der Tat wird eine Suche nach exakter Übereinstimmung "über" als Präposition anzeigen. Es kann davon ausgegangen werden, dass das Präfix in der trennbaren Variante häufig am Ende des Satzes steht, d. h. danach folgt ein Punktsymbol. Diese Variante kann formalisiert werden, betrifft aber nur solche Zeitformen wie das Präsens und Präteritum. In einer untrennbaren Variante ist eine Suche mit dem regulären Ausdruck "(\b+([Üü]ber((\w)+))\b+)" möglich. Offensichtlich steht das Präfix am Anfang des Wortes, dann folgt eine bestimmte Anzahl von Zeichen. Da dies jedoch Fälle einschließen kann, in denen das Präfix trennbar ist, z. B. "übergegangen", können wir diese Suchoption nicht als verifiziert bezeichnen.

Das heißt, fundierte Kenntnisse der zu untersuchenden Sprache helfen, die Einschränkungen der Suche zu verstehen, und den Erhalt falscher Ergebnisse zu vermeiden.

Schlussfolgerungen

Wir können schlussfolgern, dass ein wirklich gutes Ergebnis einer solchen Forschung nur mit einem kombinierten Ansatz erzielt werden kann, der eine gemeinsame Arbeit eines erfahrenen Wissenschaftlers mit der speziell entwickelten Software voraussetzt. Die Aufgabe eines Experten, der Informations- und Kommunikationstechnologien anwendet, umfasst nicht nur die rationelle Verwendung vorhandener Software, sondern auch die Erarbeitung einer Forschungsstrategie, die Ermittlung des Wirksamkeitsgrades von Softwareprodukten und, bei Vorhandensein bestimmter Kompetenzen, die aktive Beteiligung direkt an der Entwicklung spezialisierter Softwareprodukte. Ein wichtiges Problem, das angesprochen werden muss, ist daher die Aufnahme professionell orientierter Programmierdisziplinen in linguistische Curricula.

Mit den von uns entwickelten Instrumenten ist es möglich, authentisches Fremdsprachenmaterial, das bestimmte lexikalische und grammatische Phänomene sowie linguoregionale Komponenten enthält, schnell zu sammeln und zu didaktisieren. Die erhaltenen authentischen Aussagen sind effizient in der Ausbildung zukünftiger Übersetzer und Dolmetscher aus einer Fremdsprache ins Russische, auch unter Einbeziehung von Dialektismen und territorialen Realien. Eine vielversprechende Lösung ist unserer Meinung nach auch die Integration derartiger Übungen in die Lehrveranstaltungen in Kultur der Fremdsprachenkommunikation, in denen das betrachtete Kunstwerk als Material zur Vermittlung der Grundlagen der Interpretation eines literarischen Textes, des Leseverstehens im Allgemeinen, der Grammatik als eines wichtigen Aspekts der Sprache und von Übersetzen und Dolmetschen als Mittel zur Förderung der interkulturellen Kommunikation fungiert.

Literaturverzeichnis

Bäckermann, L. (2018). Unknown city paths beyond Google, but how? [Unbekannte Pfade der Stadt jenseits von Google, aber wie?]. In *Smart City – Critical Perspectives on Digitization in Cities* [*Smart City – Kritische Perspektiven auf die Digitalisierung in Städten*]. 275-281. ISBN: 978-3-8376-4336-7

Barkovich, A.A. (2016). Corpus Linguistics: The Specifics of Modern Meta-Descriptions of Language. In *Tomsk State University Journal*. 406, 5-13. DOI: 10.17223/15617793/406/1

Caton, J.N. (2020). Using Linguistic Corpora as a Philosophical Tool. In *Metaphilosophy*. 51(1), 51-70. DOI: 10.1111/meta.12405

Davies, C., Lingwood, J., Arunachalam, S. (2020). Adjective forms and functions in British English child-directed speech. In *Journal of Child Language*. 47(1), 159-185. DOI: 10.1017/S0305000919000242

Duque, E. (2020). Neuter Pronoun Ello And Discourse Verbs in Spanish. In *Journal of Pragmatics*. 155, 273-285. DOI: 10.1016/j.pragma.2019.09.006

Giampieri, P. (2020). Volcanic Experiences: Comparing Non-Corpus-Based Translations with Corpus-Based Translations in Translation Training. In *Perspectives-Studies in Translation Theory and Practice*. Early Access. DOI: 10.1080/0907676X.2019.1705361

Gorozhanov, A.I. (2019). *Institutional Educational Virtual Environment for Linguistic Purposes: Theory and Practice***.** Kazan, Buk, 184 p. ISBN 978-5-00118-322-8

Gorozhanov, A.I., Kosichenko, E.F., Guseynova, I.A. (2018). Teaching Written Translation Online: Theoretical Model, Software Development, Interim Results. In *SHS Web of Conf. (CILDIAH-2018)*. 50, 1-6. DOI: 10.1051/shsconf/20185001062

Khomenko, A.Yu. (2019). Linguistic Attributional Examination of Short Written Texts: Qualitative and Quantitative Methods. In *Political Linguistics*. 2(74), 177-187. DOI: 10.26170/pl19-02-20

Meidani, M. (2019). *Persian Calligraphy: A Corpus Study of Letterforms*. Routledge, 330 p. ISBN: 978-0-42955-907-5

Mordovin, A.Yu. (2013). Philosophical and Theoretical Background on Development of Text Corpora. In *Journal of Siberian Federal University. Humanities & Social Sciences*. 1(6), 50-61.

Potapova R., Potapov V. (2015). Cognitive Mechanism of Semantic Content Decoding of Spoken Discourse in Noise. In *Speech and Computer. SPECOM 2015. Lecture Notes in Computer Science*. 9319, 153-160. DOI: 10.1007/978-3-319-23132-7_19

Ryndina, O.M. (2019). Ethnic Culture, Digitalization and Neo-Traditionalism. In *Tomsk State University Journal of Cultural Studies and Art History*, 35, Pages: 264-274. DOI: 10.17223/22220836/35/24

Schäfer, A., Wimmer, M. (2004). *Tradition and contingency* [*Tradition und Kontingenz*]. München, Waxmann Verlag, 224 p. ISBN: 978-3-8309-6392-9

Silva, da R.C. (2020). Desterro Island [Ilha do Desterro]. In *A Journal of English Language Literatures in English And Cultural Studies*. 73(1), 129-152. DOI: 10.5007/2175-8026.2020v73n1p129

Корпусный анализ конституентов грамматических категорий текста художественного произведения с учетом лингворегионального компонента

А.И. Горожанов, И.А. Гусейнова
Московский государственный лингвистический университет
Российская Федерация, Москва

Аннотация. Ставится проблема недостаточного уровня развития профессиональной цифровизации гуманитарных наук, в том числе языкознания. Описываются результаты многолетнего исследования по разработке авторского программного обеспечения для корпусного анализа иноязычного текста, важное место в котором занимает региональный компонент. На примере романа Франца Кафки «Замок» демонстрируется работа трех авторских компьютерных программ, включая первичную обработку текста, автоматизированную разметку малого специального (грамматического) корпуса, комбинированный поиск избранных конституентов грамматических категорий определенности / неопределенности, модальности, а также отдельных грамматических явлений по маркерам размеченного корпуса и по регулярным выражениям в рамках неразмеченного текста, учитывая региональные особенности орфографии. Делается вывод о том, что максимальная эффективность исследований такого рода может быть достигнута при активной работе эксперта со специально разработанным программным продуктом. При этом в задачу эксперта входит не только профессиональное управление программным обеспечением, но и активное участие непосредственно в его разработке. Разработанная модель анализа может быть применена для анализа категориального спектра иноязычного текста, а также для составления дидактических материалов для обучения иностранным языкам и переводу.

Ключевые слова: корпусный анализ, грамматические категории, специализированное программное обеспечение, лингворегиональный компонент, межкультурная коммуникация, цифровизация, немецкий язык, Франц Кафка, регулярные выражения.

Научная специальность: 10.00.00 – филологические науки.