

DAIVA ŠVEIKAUSKIENĖ  
Institut für litauische Sprache

ARŪNAS RIBIKAUSKAS  
Gediminas technische Universität Vilnius

VYTAUTAS ŠVEIKAUSKAS  
Institut für litauische Sprache

Wissenschaftliche Forschungsrichtungen: Grammatik,  
Computerlinguistik.

## EINE MEHRSPRACHIGE WEBSITE ZUR GRAMMATIK DER LITAUISCHEN SPRACHE

Daugiakalbė lietuvių kalbos gramatikos  
informacinė sistema internete

### ANNOTATION

Vor einem Jahr begann man am Institut für litauische Sprache an einem Projekt zu arbeiten, das zum Ziel hat, die Präsentation ausführlicher Daten zu den grammatischen Eigenschaften litauischer Wörter dem Benutzer frei zugänglich im Internet zur Verfügung zu stellen. Im März 2017 wurde eine Probeversion vorbereitet, die nur Wörter mit der Wurzel „bėg“ (vgl. das Verb „bėgti/laufen“) umfasst. Die Datenbank besteht aus rund 25.000 Datensätzen. Das Projekt ist bestrebt, die Mängel schon vorhandener Arbeiten auf dem Gebiet der Computerlinguistik zu vermeiden. Das betrifft den Umfang der Wörter, die Vielseitigkeit der grammatischen Daten, die Klarheit und Verständlichkeit der Darstellung der Daten u.a. Die grammatische Information über das vom Benutzer eingegebene Wort wird in drei Bereichen dargestellt: Struktur des Wortes, morphologische Daten und morphemische Daten. Dies ist die erste Website in Litauen, auf der die Information über Typen von Morphemen dem Benutzer angegeben wird. Für jedes in der Datenbank enthaltene Wort kann man eine Flexionstabelle bekommen. Für manche Wörter sind auch Verwendungsbeispiele angegeben. Die Website wird in sieben Sprachen geführt: Litauisch, Englisch, Deutsch, Französisch, Italienisch, Russisch und Japanisch, damit Ausländer, die sich für die litauische Sprache interessieren, sie nutzen können.

SCHLÜSSELWÖRTER: Grammatik, Morphologie, Morphemik, Informationssystem, Computerlinguistik.

## ANNOTATION

The project was started in the Institute of Lithuanian language in 2016. It aims at presenting detailed data about the grammatical features of Lithuanian words. The goal is to make those data freely accessible to any user on the Internet. The preliminary version covering the words with the root “bėg/run” was published in March 2017. The database contains around 25,000 entries. We are trying to avoid drawbacks that are specific to the works done in the field of computational linguistics. It includes word coverage, grammatical data comprehensiveness, clarity and clearness of presented information etc. Grammatical information about the word in question is presented from three aspects: word structure, morphological data, and morphemic data. It is the first website in Lithuania providing morphemic types. One can get an inflection table for each word the database contains. Usage examples are given for some words. The information on the website is available in seven languages – Lithuanian, English, German, French, Italian, Russian, and Japanese – so foreigners interested in Lithuanian language can use it as well.

KEYWORDS: Grammar, Morphology, Morphemik, Information system, Computational linguistics.

## 1. EINFÜHRUNG: COMPUTERLINGUISTIK

Bis zur Mitte des 20. Jahrhunderts wurden alle Grammatiken auf Papier gedruckt. Nach dem Aufkommen von Computern (im Jahre 1942 wurde der erste Computer der Welt, MARK I, an der Universität Harvard gebaut (Schwanke 1991: 69)) begannen sie in alle Gebiete des Lebens einzudringen. Sprachen waren keine Ausnahme: bald wurde klar, dass Computer nicht nur Zahlen, sondern auch beliebige Symbole verarbeiten können. Folglich können sie auch Sprachen verarbeiten. Mehrere Linguisten begannen formale Beschreibungen von Sprachen zu entwickeln, damit die Fähigkeiten von Computern auch auf Sprachen angewendet werden konnten (Winograd 1983: 23). Auf dem Feld der künstlichen Intelligenz wurde die Sprachverarbeitung zu einem der wichtigsten Anwendungsgebiete (Chabris 1989: 71). Seit 1968 bis 1978 war die grammatische Analyse von Sprachen das wichtigste Thema der Erforscher künstlicher Intelligenz (Nierenburg 1987: 26). Die Sprachen jedoch, die sich so leicht den Menschen fügen, erschienen als harte Nuss für die Computer (Jensen, Heidorn,

Richardson 1993: 2). Deshalb konnten bislang keine guten Ergebnisse auf dem Gebiet der Sprachverarbeitung mit Hilfe von Computern erreicht werden.

### 1.1. STATISTISCHE METHODEN

Sehr große Bedeutung haben in der Sprachverarbeitung statistische Methoden. Die ersten Vorschläge, sie zu benutzen, waren kurz nach dem Aufkommen von Computern vorgebracht worden. Besonders gilt das für die englische Sprache. Dawid W. Reed erörtert die quantitative linguistische Analyse (Reed 1949: 236). Warren Weaver war der erste, der den Gedanken äußerte, Computer für Übersetzungen zu benutzen. Im Jahre 1949 schlug er vor, dazu statistische Methoden zu verwenden. Aber die damaligen Wissenschaftler lehnten es nach einer kurzen Weile ab, wahrscheinlich wegen des ungenügenden Umfangs elektronisch lesbarer Texte und wegen der nicht allzu großen Leistungsfähigkeit der damaligen Computer. Nach ein paar Jahrzehnten, als die Textkorpora gesammelt waren, und die Anwendung statistischer Methoden in der Spracherkennung gute Leistungen hervorbrachte, kehrte man wieder zu den statistischen Methoden in der Übersetzung zurück (Brown et al. 1990: 79). In Kanada gab es besonders günstige Bedingungen dafür, weil dort das ganze Material des Parlaments wegen der beiden Staatssprachen in zwei Sprachen (Englisch und Französisch) gespeichert wird. Deshalb konnte eine genügende Menge an Paralleltexten sehr schnell gesammelt werden (Al-Onaizan, Curin, Jahr 1999: 1). Die Struktur der beiden Sprachen, d.h. der englischen und der französischen, sind sehr ähnlich: die Wortfolge ist streng (an der ersten Stelle steht das Subjekt, an der zweiten Stelle das Prädikat). Die Ähnlichkeiten in der Struktur der Sprachen erlaubten, gute Leistungen bei der Übersetzung mit statistischen Methoden zu erreichen. Aber das ist mit der litauischen Sprache nicht der Fall. Die Google-Übersetzungen, die die statistische Methode benutzen, liefern bis jetzt keine guten Übersetzungen in die litauische Sprache. Laut den Angaben des Jahres 2017 sind die Ergebnisse sowohl der Übersetzung Englisch-Litauisch als auch der Übersetzung Litauisch-Englisch schlechter als diejenigen für Englisch-Lettisch, Lettisch-Englisch, Englisch-Estnisch und Estnisch-Englisch (Skadiņš 2017: 22).

Die statistischen Methoden drangen auch in andere Gebiete der Linguistik ein. Die ersten Arbeiten betrafen die Phonetik (Zipf 1929). 1944 wurden statistische Untersuchungen der Literaturlexik durchgeführt, d.h. wieviel Wörter wurden einmal gebraucht, wieviel Wörter zweimal, dreimal usw. (Yule 1944: 9).

Seit 1973 wird die Benennung *Dialektometrie* für die Bezeichnung des Gebietes der Linguistik benutzt, das sich mit der quantitativen Erforschung der Sprachen und Mundarten befasst (Köhler, Altmann, Piotrowski 2005: 498).

Ebenso wurden statistische Methoden auf dem Gebiet der Morphologie angewendet. Goldsmith beschreibt den Algorithmus für die Feststellung der Morpheme im Wort so: „algorithm that takes as input a list of words and provides as output a segmentation of the words into morphemes“ (Goldsmith 2010: 365).

Ende des vorigen Jahrhunderts werden Versuche in der russischen Literatur beschrieben, eine automatische syntaktische Analyse von Sätzen mit Hilfe statistischer Berechnungen durchzuführen. Die Wissenschaftler in Russland haben vorgeschlagen, die syntaktische Struktur von Sätzen nicht auf Grund der linguistischen Analyse zu machen, sondern eine statistische Verarbeitung der Texte zu benutzen. Sie behaupteten, dass jede sprachliche Einheit (Wort, syntaktische Kategorie des Wortes, syntaktische Konstruktion) mit einer gewissen Häufigkeit im Text vorkommt. Z.B. laut den Angaben des englischen Frequenzwörterbuchs der Valenz sind 195 verschiedene Muster der syntaktischen Beziehungen dem Verb eigen. Doch allein die zehn am häufigsten vorkommenden Muster machen 86% aller in Texten verwendeten Fällen aus. Statistische Gesetzmäßigkeiten werden auch für die lineare Struktur des Satzes benutzt. Jede Wortklasse hat ihre Frequenz, mit der sie mit einem gewissen Abstand vom Ausgangspunkt entfernt sein kann. Das Verb der englischen Sprache kann mit der Wahrscheinlichkeit 0,7 in der zweiten bis fünften Position vom Anfang des Satzes stehen. Und mit der Wahrscheinlichkeit 0,95 kommt es an der zweiten bis zehnten Position vor (Сердюков 1979: 124).

Später wurden Daten für die syntaktische Analyse aus einem Textkorpus benutzt (Köhler 2012: 31).

Doch gute Resultate sind vorläufig nur bei der Verarbeitung der englischen Sprache erzielt worden. Vidas Daudaravičius, der auf dem Gebiet der Computerisierung der litauischen Syntax gearbeitet hat, behauptet: „Naivu manyti, kad metodai, kurie sėkmingai taikomi anglų kalbai, tinka ir kitoms kalboms.“<sup>1</sup> (Daudaravičius 2012: 3).

Mit Hilfe der statistischen Methoden kann man schnell rasch arbeitende Systeme schaffen, aber unter einer Bedingung: Man muss eine bestimmte Zahl an Fehlern tolerieren (Link 1). Deshalb wurde am Institut für litauische Sprache beschlossen, sich möglichst weniger auf Statistik zu berufen, weil man nach größter Genauigkeit und Zuverlässigkeit der Daten strebt.

---

<sup>1</sup> „Es ist naiv zu denken, dass die Methoden, die für die englische Sprache erfolgreich benutzt werden, auch für die anderen Sprachen ebenso passen.“

## 1.2. DIE LAGE IN LITAUEN

Viele Arbeiten sind auf dem Gebiet der Computerisierung der Litauischen Grammatik bereits erfolgt. Die meisten von ihnen sind nur für Fachleute der Computerlinguistik verständlich. Einige sind auch für die breite Öffentlichkeit bestimmt. Sie widerspiegeln aber nur einzelne Eigenschaften und Merkmale der Wörter und der Grammatik. Z.B. wurde in Kaunas an der Universität Vytautas' des Großen eine Datenbank der Morphemik auf Grundlage eines Textkorpus erstellt. Dort fehlen viele Wortformen und man benutzt sehr viele Abkürzungen, die nicht allen verständlich sind. Die Morpheme werden mit Bindestrich getrennt und keine Daten über den Typ des Morphems angegeben. Am Institut für Mathematik und Informatik in Vilnius wurde eine Datenbank für Derivation und Morphemik geschaffen. Dort ist der Typ des Morphems angegeben sowie die Grundwörter und Bestimmungswörter (Murmulaitytė 2012: 96). Leider ist die Datenbank nicht frei zugänglich. Deshalb wurde beschlossen, ein umfassendes Informationssystem zu schaffen, das für die breite Öffentlichkeit bestimmt sein und dem Benutzer ausführliche Daten bieten können soll.

Man kann zwei Extreme in der Darstellung der litauischen grammatischen Information im Internet beobachten. Das sind: das Fehlen sogar der geläufigen Wortformen und die Darstellung überflüssiger, selbst dem Muttersprachler unverständlicher Wörter. Im Informationssystem für die litauische Grammatik ist man bestrebt, diesen beiden Extreme zu entgehen. Alle Wortformen werden automatisch mit dem Computer ausgehend vom Lemma des Wortes generiert, so erhält man den ganzen Satz an Flexionsformen. Alle möglichen Ableitungen und Zusammensetzungen werden in die Datenbank eingetragen. Also gibt es kein Fehlen von Wortformen mehr. Dann werden alle vom Computer gebildeten Wörter und Wortformen von Hand geprüft und nicht existierenden Varianten weggestrichen. So erreicht man den Ausschluss überflüssiger Wörter.

## 2. DIE ARBEITEN ZUR COMPUTERLINGUISTIK IN LITAUEN

In Litauen kann man zwei Methoden der Computerbearbeitung von Sprache unterscheiden. In Kaunas werden am Zentrum für Computerlinguistik die Arbeiten aufgrund eines Textkorpus durchgeführt. In Vilnius, an der Universität und am Institut für litauische Sprache, geht man mehr von den Eigenschaften der litauischen Sprache selbst aus. Die beiden Methoden haben sowohl Vorteile als auch Nachteile. Im Weiteren wird das an einigen Beispielen gezeigt.

## 2.1. TEXTKORPUS-BASIERTE ARBEITEN

An der Universität Vytautas' des Großen in Kaunas wurde ein Textkorpus der litauischen Gegenwartssprache erstellt. Man benutzt es für verschiedene Arten der Sprachverarbeitung. Auf dem Gebiet der Grammatik wurde ein Wörterbuch für Morphemik (Rimkutė, Kazlauskienė, Raškinis 2011) und nach einigen Jahren auch eine Datenbank geschaffen. Dem Benutzer werden sowohl die morphemischen als auch die morphologischen Daten geboten. Dies ist die erste frei zugängliche Quelle, in der das Wort in seine Morpheme zerteilt wird. Der Typ des Morphems wird leider nicht angegeben und manchmal führt das zu einer verwirrenden Darstellung der morphemischen Struktur des Wortes. Das ist der Fall, wenn Wörter mit verschiedener Struktur gleich dargestellt werden. Ein Beispiel dafür wird in Bild 1 gezeigt. Die Wörter „per-ei-ti“ (übergehen) und „per-ė-ti“ (brüten) unterscheiden sich äußerlich kaum. Die morphemischen Strukturen der beiden Wörter sehen in der Datenbank für Morphemik gleich aus. Im ersten Wort ist aber das Morphem „per“ das Präfix und im zweiten Wort ist dasselbe erste Morphem „per“ die Wurzel.

|           |                       |
|-----------|-----------------------|
| per-ei-ti | pereiti; vksm. bendr. |
| per-ė-ti  | perėti; vksm. bendr.  |

BILD 1. Wörter mit verschiedener morphemischer Struktur sind gleich dargestellt (Rimkutė, Kazlauskienė, Raškinis 2011: 8, 35)

Als den zweiten Mangel kann man das Fehlen sogar geläufiger Wortformen erwähnen. Die Partizipien haben in der litauischen Sprache sechs Kasus. In der Tabelle 1 wird das Vorkommen der Wortformen (Singular, maskulin) des Partizips „bėgantis / laufender, fließender“ in der Datenbank für Morphemik gezeigt.

Es zeigt sich, dass sogar das Lemma (Nominativ, Singular) dieses Worts fehlt. Auch die anderen fehlenden Formen, z.B. der Lokativ (*bėgančiam vandenyje* / *im fließenden Wasser*) kann man nicht als selten vorkommend oder ungeläufig betrachten. Also sind nur 2 Wortformen von 6 Kasus vorhanden.

Die Wissenschaftler der zweiten baltischen Sprache, des Lettischen, sprechen auch von einer unzureichenden Zahl an Wortformen im Textkorpus (Paiķens, Rituma, Pretkalniņa 2013: 272).

TABELLE 1. Das Vorkommen in der Datenbank (Link 2) der Wortformen des Partizips *bėgantis* / *laufender*, *fließender* im Singular, maskulin

| Kasus        | Wort       | Datenbank |
|--------------|------------|-----------|
| Nominativ    | bėgantis   | fehlt     |
| Genitiv      | bėgančio   | vorhanden |
| Dativ        | bėgančiam  | fehlt     |
| Akkusativ    | bėgantį    | vorhanden |
| Instrumental | bėgančiu   | fehlt     |
| Lokativ      | bėgančiame | fehlt     |
| Insgesamt:   | 6          | 2         |

## 2.2. AUSGANGSPUNKT – DIE LITAUISCHE SPRACHE

An der Universität Vilnius geht man von den Eigenschaften der litauischen Sprache aus. Die Website Morfologija.lt (Link 3) stellt sogar die Flexionstabellen des Wortes dar. Leider gibt es viele leere Stellen, manchmal für das ganze Paradigma. Unbekannte und für einen Litauer unverständliche Wörter kommen manchmal auch vor, z.B. „susitikimoji“ (Link 4) und ähnliche.

Am Institut für Mathematik und Informatik in Vilnius wurde eine Datenbank für Wortbildung und Morphemik geschaffen (Murmulaitytė 2012). Darin ist sehr wertvolle Information enthalten (der Typ des Morphems, Grundwort, Bestimmungswort u.a.), aber die Daten sind leider nicht frei zugänglich.

## 3. DAS INFORMATIONSSYSTEM FÜR DIE LITAUISCHE GRAMMATIK

Die litauische Sprache gehört zur Gruppe der am wenigsten computerisierten Sprachen Europas (Vaišnienė, Zabarskaitė 2012: 35). Deshalb wurde beschlossen, ein Informationssystem zu schaffen, das ausführliche Daten über die grammatischen Eigenschaften der litauischen Wörter enthält.

Die Notwendigkeit, mehr Software für einen breiten Benutzerkreis zu schaffen, wird von den Wissenschaftlern, die sich mit den baltischen Sprachen befassen, oft betont. In den Anweisungen der Konferenz BSNLP (Balto-Slavic Natural Language Processing) 2015 sowie 2017 schreibt man: „submissions

describing systems, that are made available to the wider public would be strongly encouraged / besonders gefördert werden die Einreichungen, die der breiten Öffentlichkeit zur Verfügung gestellte Systeme beschreiben“ (Link 5).

Deswegen kam man auf den Gedanken, das Informationssystem für die litauische Grammatik für einen breiten Benutzerkreis zu konzipieren und die Daten möglichst einfach und verständlich darzustellen.

Demzufolge wurde ein responsives Webdesign (RWD) gewählt, um den Bedürfnissen der Benutzer möglichst gut entsprechen zu können. Folglich wird die grammatische Information über das Wort in die Kacheln (englisch *tiles*) ausgelegt (Link 6). Deshalb kann auf die Website auch mit dem Mobiltelefon zugegriffen werden.

### 3.1. METHODE DER ERSTELLUNG DER DATENBANK

Ziel der Computerisierung der Grammatik ist, die ganze grammatische Information über jedes Wort und alle seine Formen in elektronischer Form zu haben. Es gibt zwei Wege, das zu erreichen. Der erste Weg ist, eine formale Beschreibung der grammatischen Regeln zu schaffen (das wäre eine Art Hilfsmittel) und danach den Computer die notwendigen Angaben generieren zu lassen. Der zweite Weg ist, die ganze grammatische Information über alle Formen aller Wörter im Computer zu speichern und diese bei Bedarf auszulesen. Darüber, welches Vorgehen mehr Arbeit braucht, ist noch zu diskutieren. Am Institut für Mathematik und Informatik ist die Software für eine morphologische Analyse der litauischen Sprache (Zinkevičius 2000) entwickelt worden, wozu die erste Methode benutzt worden war. Es gelang aber nicht, eine 100%ige Genauigkeit zu erreichen: Es gibt fehlerhafte Angaben, manche litauische Wörter werden nicht erkannt.

Bei der Entwicklung des Informationssystems für die litauische Grammatik wurde auf diesen Weg aus folgendem Grund verzichtet: Damit der Computer selbst die Wörter verarbeiten kann, muss man ihm sehr genau anzeigen, welche Tätigkeiten er mit welchem Wort machen soll. Also müssen zuerst alle Wörter auf solche Weise in Gruppen geteilt sein, dass für alle Mitglieder einer Gruppe dieselben Regelmäßigkeiten gelten, d.h. damit dieselbe Methode der Verarbeitung benutzt werden kann. Um das Wort in die eine oder andere Gruppe einzutragen, muss man viele Merkmale feststellen, aufgrund deren die Wörter gruppiert werden sollen. Z.B. benutzt man in der syntaktischen Analyse der litauischen Sprache das semantische Merkmal [Zeit] für die Bestimmung der syntaktischen Funktion von Akkusativ der Substantive. Im Satz „*Visą naktį ji*

skaitė šią knygą / Die ganze Nacht hat sie dieses Buch gelesen.“ und „Visą knygą ji perskaitė šią naktį / Das ganze Buch hat sie in dieser Nacht gelesen.“ Die beiden Wörter *knyga/Buch* und *naktį/Nacht* unterscheiden sich hinsichtlich der morphologischen Kategorien nicht – beide sind feminin, Singular, Akkusativ und nur das semantische Merkmal [Zeit], das das Wort *naktis/Nacht* hat, und das Wort *knyga/Buch* nicht hat, erlaubt dem Computer richtig das Wort *naktį/Nacht* als Zeitbestimmung und das Wort *knyga/Buch* als Objekt im Satz betrachten. Etwas Ähnliches muss man auch für Morphologie machen. Einer der Fehler in der schon geschaffenen Software zur morphologischen Analyse der litauischen Sprache besteht darin, dass das Morphem *-ėj-* als Suffix bei solchen Wurzeln betrachtet wird, mit denen es nicht in eine Verbindung treten darf. Man muss die Merkmale für Wörter bestimmen, die den Computer entscheiden lassen, mit welchen Wurzeln dieses Suffix *-ėj-* in Verbindung treten darf. Aber es ist sehr kompliziert das zu tun. Welches Merkmal ist z.B. den Wörtern *gerovė/Wohlhaben*, *žinovas/Kenner(Expert)* und *daržovė/Obst* gemeinsam? Dass sie alle das Suffix *-ov-* haben. Und was unterscheidet sie von allen übrigen Wörtern, an deren Wurzel sich dieses Suffix nicht fügen kann (man kann nicht sagen *\*kėdovė* – die Ableitung mit dem Suffix *-ov-* vom Wort *kėdė/Stuhl*)? Und welches Merkmal zeigt das?

Also wurde bei der Entwicklung des Informationssystems für die litauische Grammatik der zweite Weg gewählt, d.h. es wurde eine Datenbank vorbereitet, die ausführliche grammatische Informationen über alle Formen aller Wörter der litauischen Sprache besitzt. Hier muss man nicht von den grammatischen Kategorien ausgehen (wie es in allen auf Papier gedruckten Grammatiken der Fall ist), sondern von dem Wort selbst: für jedes Wort muss die ganze mit ihm verbundene Information dem Benutzer bereitgestellt werden.

Es wurde beschlossen, als Anfang eine Wurzel zu nehmen und eine Software zu entwickeln, die die ganze Struktur der litauischen Grammatik widerspiegeln soll. Als erste Probe wurde die Wurzel *bėg* (vom Verb *bėgti/laufen*) gewählt. Dann wurden alle möglichen Ableitungen mit dem Computer generiert und später alle in der litauischen Sprache nicht existierenden Wörter von Hand gestrichen. Insgesamt enthält die Datenbank rund 25.000 Einträge.

### 3.2. BESONDERHEITEN VON „LIGIS“

Das Informationssystem für die Grammatik der litauischen Sprache (**Lietuvių kalbos gramatikos informacinė sistema – LIGIS**) (Link 7) hat zwei wesentliche Merkmale. Erstens gibt es dem Benutzer mehr ausführliche Information über ein konkretes Wort als die auf Papier gedruckten Grammatiken. Und

zweitens umfasst es mehr Wörter als die Wörterbücher. In Grammatiken werden die Regeln angegeben, die für eine bestimmte Gruppe von Wörtern passen. Aber dort werden nicht alle Wörter, für die diese Regel gilt, erwähnt. LIGIS sammelt die ganze mit dem vom Benutzer eingegebenen Wort verbundene Information und gibt sie auf einer Website an. In Wörterbüchern sind nicht alle Wörter der litauischen Sprache eingetragen, dort fehlen viele Ableitungen und Komposita. LIGIS umfasst alle möglichen Ableitungen. Zum Vergleich kann ein Beispiel gegeben werden. Die Datenbank für Morphemik hat rund 75.000 Einträge. Aber sie umschreiben verschiedene litauische Wörter. Die LIGIS-Datenbank besteht momentan aus 25.000 Einträgen und umfasst nur die Wörter von einer einzigen Wurzel: *bėg-*. In der Datenbank für Morphemik gibt es 118 Wörter mit dieser Wurzel.

### 3.3. STRUKTUR DES WORTES

Nach der Eingabe des Wortes wird dem Benutzer die Wortstruktur erläutert: Das Lemma und Grundwort sowie das Bestimmungswort für Ableitungen und Zusammensetzungen werden angegeben. Wenn das Wort grammatisch mehrdeutig ist, so bekommt jede der Bedeutungen schon eine Nummer für das Lemma. Und später werden die morphologischen und morphemischen Daten unter dieser Nummer angezeigt. Ein Beispiel für ein mehrdeutiges Wort wird im Anhang A angeführt.

**Übrige Formen.** So heißt die Schaltfläche (englisch *button*), die sich in der Kachel der Struktur des Wortes befindet. Die Website der morphologischen Analyse der russischen Sprache (Link 8) stellt alle Wortformen für das eingegebene Wort jedes Mal dar. Im Informationssystem für litauische Grammatik wurde beschlossen, eine Schaltfläche einzuführen und über den Link die übrigen Formen vom Benutzer abrufen zu lassen, weil der Benutzer nicht jedes Mal alle Formen eines Wortes braucht.

### 3.4. MORPHOLOGISCHE DATEN

Die morphologische Information umfasst die Benennung der Wortart und die grammatischen Kategorien: Kasus, Genus, Zahl für Substantive; Person, Tempus, Modus für Verben usw. Die mehrdeutigen Wörter bekommen eine Nummerierung, wenn z.B. ein Substantiv einige homonyme Kasusformen hat.

In der Datenbank für Morphemik (Link 2) gibt es auch die morphologischen Daten über das Wort. Aber die Information wird in lauter Abkürzungen

angegeben. Ein solches Beispiel wird im Bild 2 angeführt. Für Nichtfachleute können so viele Abkürzungen oft unverständlich sein. Besonders für Ausländer, die die litauische Sprache studieren oder lernen, können auf diese Weise dargestellte Informationen bisweilen unklar sein.

|                        |                                                |
|------------------------|------------------------------------------------|
| <b>bėg-anč-io</b>      |                                                |
| Žodžio lema:           | <b>bėgti</b>                                   |
| Dažnumas:              | <b>1</b>                                       |
| Gramatinė informacija: | <b>dlv. vyr. g. es. l. vns. kilm. veik. r.</b> |

BILD 2. Daten über das Wort *bėgančio*, dargestellt in lauter Abkürzungen (Link 2)

Im Informationssystem für litauische Grammatik werden keine Abkürzungen benutzt. Alle Angaben werden ausgeschrieben (Anhang A).

### 3.5. MORPHEMISCHE DATEN

Der wichtigste Unterschied zu den Datenbanken, auf die bereits heute zugegriffen werden kann, besteht in der Angabe des Typs für jedes Morphem. Nur auf diese Weise erreicht man Eindeutigkeit bei der Darstellung der Morpheme eines Wortes. Sonst kann man etwa die zweite Wurzel in einem zusammengesetzten Wort nicht vom Suffix unterscheiden, was in der Datenbank für Morphematik der Fall ist.

Das erste und entscheidende Merkmal war die Darstellung der Morpheme in verschiedenen Farben. Für jeden Morphemtyp (Wurzel, Präfix, Suffix, Endung) wurde eine Farbe bestimmt. Präfixe werden blau geschrieben. Für die Wurzel wurde Rot benutzt. Suffixe werden grün dargestellt und für die Endung wurde Schwarz gewählt.

Zusätzlich wird das in Morpheme zerlegte Wort in vertikaler Richtung ausgelegt und daneben wird nicht nur die Benennung des Typs jedes Morphems des Wortes geschrieben, sondern es werden auch die Eigenschaften des Morphems angegeben, falls das Morphem solche besitzt, z.B. für ein Suffix: Wortbildungs- oder Formbildungs-; für Endungen: verkürzte, pronominalisierte u.ä. Im Fall der verkürzten Endungen, die in der gesprochenen Sprache sehr verbreitet sind, wird daneben auch die volle Endung angezeigt. Als Beispiel kann man die Worte *in der gesprochenen Sprache* anführen: *šnekamojoje kalboje* – volle Endungen, *šnekamoj kalboj* – verkürzte Endungen. Hat das Wort die volle Endung, so wird kein Merkmal der Länge angegeben, weil die meisten Wörter die vollen Endungen haben, und eine solche Information überflüssig wäre. Wenn

die Endung nicht pronomialisiert ist, wird die Information darüber auch nicht angegeben.

Lautveränderungen werden als eine Eigenschaft der Wurzel betrachtet. Es können sowohl Konsonanten- als auch Vokalwandel sein. Noch eine weitere Art der Lautveränderungen kann man in der litauischen Sprache beobachten: den Schwund von Konsonanten im Imperativ. Und auch Infixe gehören zu den Lautveränderungen in der Wurzel eines Wortes. Für eine Eigenschaft des Präfixes wird seine Herkunft gehalten: präpositionalen Ursprungs, aus einer Partikel entstanden oder internationalen Ursprungs.

Im Fall der grammatischen Mehrdeutigkeit von Wörtern werden die morphemischen Strukturen geprüft. Wenn sie gleich sind, wird nur eine Struktur geschildert. Bestehen indes Unterschiede, so werden die morphemischen Strukturen für jede Bedeutung des Wortes angegeben (Anhang A).

### 3.6. ALLE FORMEN DES WORTES

Für jedes in der Datenbank enthaltene Wort kann man die Flexionstabelle über einen Klick auf die Schaltfläche „Übrige Formen“ abrufen. Dort sind alle Daten aus der morphologischen Kachel wiedergegeben. Oben in der Tabelle werden die Kategorien angegeben, die die Flexionstabelle nicht enthält, z.B. werden die Substantive in der litauischen Sprache nicht nach Genus flektiert, somit wird das Genus oben in der Tabelle angegeben sowie das Lemma und die Wortart. Die Adjektive und Partizipien haben drei Genera und eine Besonderheit der litauischen Sprache besteht darin, dass nur zwei davon (maskuline und feminine) alle sechs Kasus besitzen. Das dritte Genus hat nur eine Form, weshalb sie ohne Kasusbenennung angegeben wird (Anhang B). Für Verben werden alle Tempusformen in allen Modi dargestellt (Anhang D).

Die Konjugationsklassen sowie die Deklinationsklassen werden auf diese Weise entbehrlich, deshalb werden dazu keine Daten angegeben.

Für manche Wörter sind auch Verwendungsbeispiele in die Flexionstabelle eingetragen. Dazu wird Schnellinfo (englisch *tooltip*) benutzt. Im Anhang C ist als Beispiel für das Wort „bėgi“ (Akkusativ, Singular vom Substantiv *bėgis* / *Eisenbahnschiene, Lauf*) angeführt.

### 3.7. DIE DATENBANK

Bei der Entwicklung der Datenbank strebte man nach möglichst hoher Qualität und Zuverlässigkeit. Deshalb wurde ein großer Teil der Arbeit per Hand

gemacht, weil nur mit Hilfe des Menschen eine maximal zuverlässige Information erreicht werden kann (Sulger 2013: 553).

Am Institut für litauische Sprache wurden die Forschungen zu den litauischen Präfixen durchgeführt. Es wurde festgestellt, dass es in der litauischen Sprache rund 600 Kombinationen von Präfixen gibt (Šveikauskienė 2015: 195). Unter „Kombinationen“ werden verschiedene Sätze von Präfixen verstanden, die vor der Wurzel erscheinen können z.B. *at-bėgti*, *nu-bėgti*, *ne-be-at-bėgti*, *ne-nu-bėgti*, *te-be-bėgti* usw. Es zeigte sich, dass nur 220 von diesen Kombinationen mit den Wurzeln von Verben gebraucht werden; die anderen, z.B. *nuo-*, werden mit Substantiven gebraucht: *nuo-monė* /die Meinung/. Die Software wurde so ausgelegt, dass sie die Wurzel automatisch mit allen möglichen Kombinationen von Präfixen verbindet. Danach werden die vom Computer gegebenen Resultate von Menschen überprüft und die in der litauischen Sprache tatsächlich vorhandenen Wörter in die Datenbank eingetragen. So gelingt es, zwei Extreme zu vermeiden: Die übermäßige Anzahl der vom Computer generierten Formen wird ebenso vermieden wie eine Unzulänglichkeit des Umfangs der Wörter.

Alle Lemmata und die sie betreffende grammatische Information wird von Hand in die Datenbank eingetragen. Da die am Institut für Mathematik und Informatik geschaffene Software bei der Generierung der Wortformen zu dem gegebenen Lemma keine Fehler macht, wurde dem Computer zugetraut, die Einträge der übrigen Wortformen automatisch zu bilden.

Bemerkenswert ist, dass in LIGIS auch die Wörter mit verkürzten Endungen bearbeitet werden. Solche Wörter kommen in der gesprochenen Sprache sehr häufig vor und keine der schon vorhandenen Datenbanken bearbeitet diese Wortformen. Dazu muss betont werden, dass LIGIS auch solche Ableitungen enthält, die sogar im Wörterbuch der litauischen Sprache in 20 Bänden sowie seiner elektronischen Version (Link 9) fehlen, z.B. *nebeužbėgti/nicht mehr hinauflaufen*.

In der Datenbank wird die Information in einer computerfreundlichen Form aufbewahrt, d.h. es wird die Kodierung in Prague Markup Language (Hana, Štěpánek 2012) benutzt. Diese Form der Daten wird im Internet nicht vorgestellt, weil es für Nichtprofessionals der Computerlinguistik überflüssig wäre.

Dass die ganze Information in einer Datenbank aufbewahrt wird, macht LIGIS auch für die Sprachwissenschaftler, die die litauische Sprache erforschen, zu einem nützlichen Werkzeug. Man kann Auskunft verschiedenster Art bekommen, z.B. kann man Wörter mit einer bestimmten Kombination von Präfixen und Suffixen u.a. suchen.

### 3.8. MEHRSPRACHIGKEIT DER WEBSITE

Die Website wird in sieben Sprachen geführt: Litauisch, Englisch, Deutsch, Französisch, Italienisch, Russisch und Japanisch. Die Notwendigkeit, eine mehrsprachige Website zu haben, wurde von den nicht befriedigenden litauischen Google-Übersetzungen verursacht. Die maschinelle Übersetzung der litauischen Sprache ist unzuverlässig. Deshalb wurde beschlossen, die Website für die litauische Grammatik mehrsprachig zu gestalten und nur von Menschen übersetzte Texte zu benutzen. Die Lage für andere Sprachen ist offensichtlich nicht viel besser. Davon zeugt der Brief von LinguistList vom 8. Juni 2017, in dem die Bitte geäußert wird, die Übersetzung der Benennung LINGUIST List von Muttersprachlern geben zu lassen und keinesfalls eine maschinelle Übersetzung (LinguistList 2017: 28.2524).

Bei der Entscheidung, die Website mehrsprachig zu gestalten, war unsere Idee folgende: Wenn ein Ausländer, z.B. ein Norweger, die litauische Sprache lernt oder studiert und manche litauische Wörter für ihn bekannt sind, manche jedoch noch nicht, so kann er eine Sprache wählen, die er besser als die litauische kann, z.B. Deutsch, und alle ihm unbekannten litauischen Wörter werden ihm verständlich. Zu diesem Zweck wurde sogar die Speicherung des vom Benutzer eingegebenen Wortes bei der Umschaltung der Sprachen vorgesehen. Also wenn ein Benutzer eine andere Sprache wählt, muss er das Wort nicht neu eingeben.

## 4. PLÄNE FÜR DIE ZUKUNFT

An der Universität Vilnius liegen viele zweisprachige Wörterbücher digitalisiert vor. Am Institut für litauische Sprache wurde das große Wörterbuch in 20 Bänden digitalisiert sowie die anderen einsprachigen Wörterbücher: Synonym-, Antonym-, Phraseologismenwörterbücher u.a. Es ist geplant, die beiden Arten von Wörterbüchern mit dem Informationssystem der litauischen Grammatik zu verbinden und ein Analogon zur deutschen Website CANOONET (Link 10) zu bilden.

Da das Informationssystem für die Grammatik der litauischen Sprache die Daten über die Morphemtypen enthält, ist es möglich, eine Suche nach Wörtern aufgrund des Morphemmodells zu machen. Und eben dies ist für die Zukunft vorgesehen.

Ein Tab (englisch *tab*) ist in der Website LIGIS für Theorie bestimmt und befindet sich zurzeit in Bearbeitung. Dort sollen vertiefte akademische Kenntnisse zur litauischen Grammatik angegeben sein.

In der ersten Etappe beschäftigt man sich mit Morphologie. Auf der zweiten Stufe sollen auch syntaktische Daten eingetragen werden.

## 5. SCHLUSSFOLGERUNGEN

1. Diese Herangehensweise kann auch für andere stark flektierte Sprachen nützlich und sinnvoll sein.
2. Das Informationssystem für litauische Grammatik kann für Schüler, Studenten, für Ausländer, die Litauisch lernen, von großem Nutzen sein, ebenso für Sprachwissenschaftler, die die litauische Sprache erforschen.
3. Die gesammelten Daten können auch als Dokumentation der litauischen Sprache dienen.

## LITERATURVERZEICHNIS

Al – Onaizan Yaser, Curin Jan, Jahr Michael, Knight Kevin, Lafferty John, Melamed Dan, Och Franz –Josef, Purdy David, Smith Noah A., Yarowsky David 1999: Statistical Machine Translation. – *Final Report* JHU Workshop. <http://mt-archive.info/JHU-1999-AlOnaizan.pdf> (Zugriff am 21.11. 2017).

Brown Peter F., Cocke John, Della Pietra Stephen A., Della Pietra Vincent J., Jelinek Frederick, Lafferty John D., Mercer Robert L., Roossin Paul S. 1990: A Statistical Approach to Machine Translation. – *Computational Linguistics* Volume 16, Number 2, June, 79–85. <http://www.aclweb.org/anthology/J90-2002> (Zugriff am 21.11. 2017).

Chabris Christopher F. 1989: *Artificial Intelligence & Turbo C*. Homewood: Dow Jones-Irwin.

Daudaravičius Vidas 2012: *Teksto skaidymas pastoviųjų junginių segmentais*. Doktoro disertacijos santrauka. Kaunas: Vytauto Didžiojo universitetas.

Goldsmith John A. 2010: Segmentation and Morphology. *The Handbook of Computational Linguistics and Natural Language Processing*. Wiley-Blackwell, 364–393.

Hana Jirka, Štěpánek Jan 2012: Prague Markup Language Framework. *Proceedings of the 6th Linguistic Annotation Workshop*. Jeju, Republic of Korea, 12–13 July, 12–21.

Jensen Karen, Heidorn George Emil, Richardson Stephen D. 1993: *Natural language processing: The PLNLP Approach*. Boston / London: Kluwer Academic Publishers.

Köhler Reinhard 2012: *Quantitative Syntax Analysis*. De Gruyter Mouton.

Köhler Reinhard, Altmann Gabriel, Piotrowski Rajmund G. 2005: *Quantitative Linguistik*. Berlin / New York: Walter de Gruyter.

Linguist List 2017: 28.2524, Qs: How do you say the LINGUIST List in your language? 08 Jun 2017. [linguist@listserv.linguistlist.org](mailto:linguist@listserv.linguistlist.org).

Murmulaitytė Daiva 2012: Lietuvių kalbos morfemikos ir žodžių darybos tyrimų perspektyvos. – *Žmogus ir žodis* 1(14), 96–102.

Nirenburg Sergei 1987: *Machine Translation: Theoretical and Methodological Issues*. London: Cambridge University Press.

Paikens Pēteris, Rituma Laura, Pretkalniņa Lauma 2013: Morphological Analysis with limited resources: Latvian example. *Proceedings of the 19th Nordic Conference of Computational Linguistics. (NODALIDA 2013)*; Linköping Electronic Conference Proceedings #85, 267–277.

Reed David W. 1949: A Statistical Approach to Quantitative Linguistic Analysis, *WORD*, 5:3, 235–247, DOI: 10.1080/00437956.1949.11659355.

Rimkutė Erika, Kazlauskienė Asta, Raškinis Gailius 2011: *Dažninis lietuvių kalbos žodynas*. Kaunas: VDU.

Schwanke Martina 1991: *Maschinelle Übersetzung: Ein Überblick über Theorie und Praxis*. Berlin: Springer-Verlag.

Skadiņš Raivis 2017: *Neural MT and other Language Technologies at TILDE*. [http://school.grammaticalframework.org/2017/slides/Raivis-Neural\\_MT\\_at\\_Tilde.pdf](http://school.grammaticalframework.org/2017/slides/Raivis-Neural_MT_at_Tilde.pdf) (zugegriffen 2017.11.21).

Sulger Sebastian, Butt Miriam, King Tracy Holloway, Meurer Paul 2013: ParGramBank: The ParGram Parallel Treebank. – *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, 550–560. <http://aclweb.org/anthology/P/P13/P13-1054.pdf> (zugegriffen 2017.11.21).

Šveikauskienė Daiva 2015: Morphemic structure of the Lithuanian prefixes. – *Language: Meaning and Form 6. – Language System and Language Use*. Rīga: Latvijas universitāte, 189–197.

Vaišnienė Daiva, Zabarskaitė Jolanta 2012: The Lithuanian Language in the Digital Age. – G. Rehm, H. Uzhkoreit *White Paper Series*. Heidelberg / New York / Dordrecht / London: Springer.

Winograd Terry 1983: *Language as a Cognitive Process 1. Syntax*. London: Addison-Wesley Publishing Company.

Yule George Udny 1944: *The Statistical Study of Literary Vocabulary*. Cambridge MA, Cambridge university press.

Zinkevičius Vytautas 2000: Lemuoklis – morfologinei analizei. – *Darbai ir dienos* 24, 245–273.

Zipf George Kingsley 1929: Relative frequency as a Determinant of Phonetic Change. – *Harward studies in classical philology* 40, 1–95.

Сердюков Пётр Иванович 1979: Оптимизация алгоритма поиска синтаксических связей. – *Международный семинар по машинному переводу*. Москва: ВЦП, 123–125.

### LINKS

Link 1: <http://www.digitalgrammars.com/> (Zugriff am 21.11. 2017)

Link 2: <http://tekstynas.vdu.lt/page.xhtml?id=morfema-db> (Zugriff am 21.11. 2017)

Link 3: <http://morfologija.lt/> (Zugriff am 21.11. 2017)

Link 4: <http://morfologija.lt/zodzio-formos/susitikimas> (Zugriff am 21.11. 2017)

Link 5: <http://bsnlp-2017.cs.helsinki.fi/cfp.html> (Zugriff am 21.11. 2017)

Link 6: [https://de.wikipedia.org/wiki/Microsoft\\_Windows\\_8#Oberfl.C3.A4che](https://de.wikipedia.org/wiki/Microsoft_Windows_8#Oberfl.C3.A4che)  
(Zugriff am 21.11. 2017)

Link 7: <http://ligis.lki.lt/> (Zugriff am 21.11. 2017)

Link 8: <http://goldlit.ru/component/slog> (Zugriff am 21.11. 2017)

Link 9: <http://www.lkz.lt/Visas.asp?zodis> (Zugriff am 21.11. 2017)

Link 10: <http://www.canoo.net/> (Zugriff am 21.11. 2017)

# ANHANG A. GRAMMATISCHE INFORMATION.

Beispiel des grammatisch mehrdeutigen Wortes *nubėgti* / *hinlaufen-hingelaufene*.  
Die grammatische Information wird in der Webseite in Kacheln ausgelegt.  
Diese Darstellungsweise ist günstig für responsives Webdesign – RWD.  
Bemerkung: Die Farben sind im Bild entfernt.

## STRUKTUR DES WORTES

LEMMA

1. nubėgti  
2. nubėgtas

ABLEITUNG VON

1. bėgti  
2. nubėgti

## MORPHOLOGISCHE DATEN

1. **Verb**  
Infinitiv

2. **Partizip II**  
Passiv, Präteritum  
männlich, Plural, Nominativ

## MORPHEMISCHE DATEN

1. nu-bėg-ti

nu

Präfix

Präpositional-

bėg

Wurzel

ti

Suffix

Formbildungs-

2. nu-bėg-t-i

nu

Präfix

Präpositional-

bėg

Wurzel

t

Suffix

Wortbildungs-

i

Endung

Straipsniai / Articles

161

ANHANG B. FORMENTABELLE FÜR  
DEKLINATION.

Als Beispiel wird die Flexion des Partizips *nubėgtas* / *hingelaufene* in allen drei Genera und allen Kasus angeführt.

| NUBĖGTAS<br>PARTIZIP II, PASSIV, PRÄTERITUM |           |            |
|---------------------------------------------|-----------|------------|
| Männlich                                    |           |            |
|                                             | Singular  | Plural     |
| Nominativ                                   | nubėgtas  | nubėgti    |
| Genitiv                                     | nubėgto   | nubėgtų    |
| Dativ                                       | nubėgtam  | nubėgtiems |
| Akkusativ                                   | nubėgtą   | nubėgtus   |
| Instrumental                                | nubėgtu   | nubėgtais  |
| Lokativ                                     | nubėgtame | nubėgtuose |
| Weiblich                                    |           |            |
|                                             | Singular  | Plural     |
| Nominativ                                   | nubėgta   | nubėgtos   |
| Genitiv                                     | nubėgtos  | nubėgtų    |
| Dativ                                       | nubėgtai  | nubėgtoms  |
| Akkusativ                                   | nubėgtą   | nubėgtas   |
| Instrumental                                | nubėgta   | nubėgtomis |
| Lokativ                                     | nubėgtoje | nubėgtose  |
| Neutrum                                     |           |            |
|                                             | nubėgta   |            |

## ANHANG C. TOOLTIP FÜR VERWENDUNGSBEISPIELE.

Als Beispiel wird die Schnellinfo für die Akkusativ-Singular-Form *bėgi* des Substantivs *bėgis* / *Lauf, Getriebe, Eisenbahnschiene* angeführt.

| <b>BĖGIS</b>                              |                                                                        |               |
|-------------------------------------------|------------------------------------------------------------------------|---------------|
| <b>SUBSTANTIV, GATTUNGSNAME, MÄNNLICH</b> |                                                                        |               |
|                                           | <b>Singular</b>                                                        | <b>Plural</b> |
| <b>Nominativ</b>                          | bėgis                                                                  | bėgiai        |
| <b>Genitiv</b>                            | bėgio                                                                  | bėgių         |
| <b>Dativ</b>                              | bėgiui                                                                 | bėgiams       |
| <b>Akkusativ</b>                          | bėgį                                                                   | bėgius        |
| <b>Instrumental</b>                       | VERWENDUNGSBEISPIELE                                                   |               |
| <b>Lokativ</b>                            | Jis įjungė <b>bėgį</b> ir nuvažiavo jos pakeidaujama kryptimi.         |               |
| <b>Vokativ</b>                            | Parduodu 5 m geležinkelio <b>bėgį</b> .                                |               |
|                                           | Paroda "Pro amžių <b>bėgį</b> " skirta poeto Sirijos Giros jubiliejui. |               |

# ANHANG D. FORMENTABELLE FÜR KONJUGATION.

Als Beispiel wird die Flexion des Verbs *bėgti* / *laufen* in allen Tempusformen und Modi angeführt.

| NUBĖGTI<br>VERB                |            |             |
|--------------------------------|------------|-------------|
| Indikativ, Präsens             |            |             |
|                                | Singular   | Plural      |
| 1. Person                      | nubėgu     | nubėgame    |
| 2. Person                      | nubėgi     | nubėgate    |
| 3. Person                      | nubėga     | nubėga      |
| Indikativ, Präteritum einmalig |            |             |
|                                | Singular   | Plural      |
| 1. Person                      | nubėgau    | nubėgome    |
| 2. Person                      | nubėgai    | nubėgote    |
| 3. Person                      | nubėgo     | nubėgo      |
| Indikativ, Präteritum iterativ |            |             |
|                                | Singular   | Plural      |
| 1. Person                      | nubėgdavau | nubėgdavome |
| 2. Person                      | nubėgdavai | nubėgdavote |
| 3. Person                      | nubėgdavo  | nubėgdavo   |
| Indikativ, Futur               |            |             |
|                                | Singular   | Plural      |
| 1. Person                      | nubėgsiu   | nubėgsime   |
| 2. Person                      | nubėgsi    | nubėgsite   |
| 3. Person                      | nubėgs     | nubėgs      |
| Konjunktiv                     |            |             |
|                                | Singular   | Plural      |
| 1. Person                      | nubėgčiau  | nubėgtume   |
| 2. Person                      | nubėgtum   | nubėgtute   |
| 3. Person                      | nubėgtų    | nubėgtų     |
| Imperativ                      |            |             |
|                                | Singular   | Plural      |
| 1. Person                      | -          | nubėkime    |
| 2. Person                      | nubėk      | nubėkite    |
| Optativ, Präsens               |            |             |
|                                | Singular   | Plural      |
| 3. Person                      | tenubėga   | tenubėga    |
| Optativ, Futur                 |            |             |
|                                | Singular   | Plural      |
| 3. Person                      | tenubėgs   | tenubėgs    |

## Daugiakalbė lietuvių kalbos gramatikos informacinė sistema internete

### SANTRAUKA

Lietuvių kalbos kompiuterizavimo darbuose galima pastebėti du pagrindinius trūkumus: trūksta kartais net ir dažnai vartojamų formų ir pateikiami pertekliniai, lietuvių kalboje neegzistuojantys žodžiai. Kuriant *Lietuvių kalbos gramatikos informacinę sistemą* (LIGIS) stengiamasi išvengti jų abiejų. Visos formos generuojamos kompiuteriu automatiškai ir vėliau gauti rezultatai peržiūrimi žmogaus, kad būtų atmesti lietuviui nesuprantami žodžiai.

*Lietuvių kalbos gramatikos informacinė sistema* apima visus darinius. Šiuo metu duomenų bazę sudaro tik šaknies *bėg-* žodžiai. Jų yra apie 25 000. Palyginimui galima pateikti tokius duomenis: morfemikos duomenų bazę sudaro yra apie 75 000 įrašų, bet jie surinkti iš visos lietuvių kalbos leksikos. Su šaknimi *bėg-* ten tėra 118 žodžių.

Kuriant *Lietuvių kalbos gramatikos informacinę sistemą* pagrindinis tikslas buvo pateikti išsamią gramatinę informaciją plačiai visuomenei. Todėl buvo stengiamasi duomenis atvaizduoti kuo aiškiau ir suprantamiau. Svetainė pritaikyta naudotis ir mobiliuosiuose telefonuose.

Vartotojas, pateikęs žodį, gauna trijų tipų informaciją apie jį: žodžio struktūra (pradinė forma bei pamatiniai žodžiai dariniams), morfologinė informacija (kalbos dalis bei jos morfologinės kategorijos – linksnis, giminė, skaičius, laikas, laipsnis ir t. t.) ir morfeminiai duomenys – žodis išskaidytas morfemomis, nurodant morfemos tipą bei požymius, jei morfema jų turi, pavyzdžiui, priesaga – darybinė / kaitybinė, galūnė – įvardžiutinė, sutrumpėjusi ir kt. Kiekviena morfema žymima vis kita spalva. Reikia pabrėžti, kad *Lietuvių kalbos gramatikos informacinė sistema* yra pirmasis tinklapis Lietuvoje, kur vartotojui pateikiama informacija apie morfemos tipą. Be to, apdorojami ir žodžiai su sutrumpėjusiomis galūnėmis, kurios labai paplitusios, ypač šnekamojoje kalboje.

Kiekvienam duomenų bazėje esančiam žodžiui vartotojas gali gauti visą kaitybinę lentelę. Kai kuriems žodžiams pateikiama vartojimo pavyzdžių.

Kuriant *Lietuvių kalbos gramatikos informacinę sistemą* siekiama kuo didesnio tikslumo ir patikimumo. Todėl visos lemos (žodžių pradinės formos – vardininkas, bendratis) suvedamos į duomenų bazę rankomis. Kaitybines formas sugeneruoti patikėta kompiuteriui, nes šiame jo darbe nebuvo pastebėta klaidų.

Svetainė pateikiama septyniomis kalbomis: lietuvių, anglų, vokiečių, prancūzų, italų, rusų ir japonų.

Ateityje planuojama *Lietuvių kalbos gramatikos informacinę sistemą* jungti su suskaitmenintais žodynais ir sukurti vokiečių tinklapio CANOONET analogą.

Įteikta 2017 m. rugsėjo 5 d.

DAIVA ŠVEIKAUSKIENĖ

*Petro Vileišio g. 5, LT-10308 Vilnius, Lietuva*  
*daiva.fmf@gmail.com*

ARŪNAS RIBIKAUSKAS

*Saulėtekio al. 11, LT-10221 Vilnius, Lietuva*  
*arunas.ribikauskas@vgtu.lt*

VYTAUTAS ŠVEIKAUSKAS

*Petro Vileišio g. 5, LT-10308 Vilnius, Lietuva*  
*coder.vs713@gmail.com*