СТАТИСТИЧЕСКИ ИЗВОДИ ЗА КОЛИЧЕСТВЕНИ ХАРАКТЕРИСТИКИ НА СЛОЖНОТО СМЕСЕНО ИЗРЕЧЕНИЕ В БЪЛГАРСКИЯ ЕЗИК

Веска Нончева[а], Петя Бъркалова[б], Александър Иванов[в], Владимир Чавдаров[г]

Пловдивски университет „Паисий Хилендарски"
ул. Цар Асен 24, 4000 Пловдив
wesnon@uni-plovdiv.bg[а], barka@uni-plovdiv.bg[б], alex9408@icloud.com[в], vlado3vsch@gmail.com[г]

***Резюме**: В статията е построен математически модел на изречението в българския език, отразяващ неговите количествени характеристики обхват, брой главни изречения, брой подчинени изречения, брой съподчинени изречения, дълбочина на изречението, брой съчинителни връзки и брой подчинителни връзки. Този модел е приложен за изследване на сложното смесено изречение в българския език. Направени са статистически изводи за обхват и дълбочина на изреченската структура, както и за честотата на типовете подчинени изречения в романа „Антихрист" на Емилиян Станев.*

***Ключови думи:** анализ на данни, български синтаксис, сложното смесено изречение, математическо моделиране на естествения език, модел на изречението*

1. Сложното смесено изречение в българския език

Естественият език, според представителите на генеративната граматика, е безкрайно множество от изречения, създаващи се от краен брой думи по краен брой правила. Първичната функция на езика е процесът на формулиране, изразяване и разбиране на мисли. Възприемането на изреченската реалност като архитертура на менталния план на „говорещия" позволява моделиране на синтактичната структура (изказа). Езиковата компетентност на човека е психологическият генератор на синтактичните конструкции.

Счита се, че езиковата компетентност на Емилиян Станев е върхова сред тези способности на носителите на българския език през 20 век. Световната организация ЮНЕСКО е признала неговия принос към световното литературно и културно наследство.

Простите изречения в едно сложно смесено изречение се свързват чрез съчинителни и подчинителни връзки. Например в сложното смесено изречение „Виждам и разбирам, че си прав." има една съчинителна и една подчинителна връзка.

Количествените характеристики на сложното смесено изречение, които наблюдаваме в това изследване са: брой прости изречения в сложното (обхват),

брой главни изречения, брой подчинени изречения, брой съподчинени изречения, брой последователно подчинени изречения (дълбочина), брой съчинителни и брой подчинителни връзки между простите изречения в рамките на сложното смесено.

Като пример за сложно смесено изречение ще разгледаме изречението „Искаше ми се /да бъда сам, /да размисля/що стана с мене,/ обаче старецът не ме оставяше с мислите ми." от романа „Антихрист" на Емилиян Станев ([5], стр. 229). Пресмятаме неговите количествени характеристики и получаваме:

- обхват 5: пет прости изречения,
- брой главни изречения 2: искаше ми се / старецът не ме оставяше с мислите ми
- брой подчинени изречения 3: да бъда сам / да размисля/ що стана с мене
- брой съподчинени изречения 2: да бъда сам, /да размисля
- дълбочина на изречението 3: *искаше ми се /да бъда сам, да размисля/ що стана с мене*
- брой съчинителни връзки в изречението 2: (първата е пауза, маркирана със запетая, и втората е съчинителният съюз *обаче*): *Искаше ми се да бъда сам, да размисля що стана с мене, обаче старецът не ме оставяше с мислите ми*.
- брой подчинителни връзки в изречението 3: два подчинителни съюза *да* (*да бъда сам, да размисля*)и една безсъюзна връзка (*що стана с мене*).

Всяко изречение може да бъде представено и със синтактично дърво със съответен обхват и дълбочина [1,2].

2. Математически модел на изречението

Изречението ще моделираме със случайния вектор $X=(X_1, X_2, X_3, X_4, X_5, X_6, X_7)$, където X_i, $i=1,\ldots,7$, са дискретни целочислени случайни величини и

X_1 е брой прости изречения (обхват),

X_2 е брой главни изречения,

X_3 е брой подчинени изречения,

X_4 е брой съподчинени изречения,

X_5 е брой последователно подчинени изречения (дълбочина),

X_6 е брой съчинителни връзки в изреченията,

X_7 е брой подчинителни връзки в изреченията.

Този модел отразява количествените характеристики на изречението. С този модел можем да опишем различни видове изречения. Например:

- ($X_1=1$, $X_2=0$, $X_3=0$, $X_4=0$, $X_5=1$, $X_6=0$, $X_7=0$) е просто изречение.
- ($X_1\geq 2$, $X_2\geq 1$, $X_3\geq 1$, $X_4\geq 1$, $X_5\geq 1$, $X_6=0$, $X_7\geq 1$) е характерно за сложното съставно изречение.

- ($X_1 \geq 3$, $X_2 \geq 1$, $X_3 \geq 1$, $X_4 \geq 2$, $X_5 \geq 2$, $X_6 \geq 1$, $X_7 \geq 1$) е характерно за сложно смесено изречение.

Обект на това изследване е сложното смесено изречение.

Ще направим статистически изводи за количествените характеристики на сложното смесено изречение в произведението „Антихрист" на Емилиян Станев.

3. Честотни наблюдения на сложното смесено изречение

Направена е извадка от 200 сложни смесени изречения, избрани по случаен начин от романа „Антихрист" на Емилиян Станев [5]. За всяко изречение са пресметнати обхват, дълбочина, брой главни изречения, брой подчинени изречения, брой съподчинени изречения, брой съчинителни и брой подчинителни връзки [4]. Получените данни са представени в таблица.

Данните ще анализираме със среда за статистически анализ на данни R. (R Core Team 2012)

4. Статистически изводи за сложното смесено изречение

Медианата на една случайна величина е мярка за нейната очаквана (средна) стойност. Нарича се още медиана на популацията, защото случайната величина е модел на популацията.

Медианата на извадката е тази стойност, за която 50% от данните са по-големи от нея или равни на нея, а останалите 50% от данните са по-малки или равни на нея. Медианата на извадката може да се изчисли от данните, които имаме.

Ако анализираме всички изречения от едно произведение, можем да пресметнем медианата на тази популация. Но това обикновено е непосилна задача.

Обикновено медианата на популацията не е известна и ние трябва да я оценим. Едно средство за оценка на медианата е доверителният интервал. Доверителният интервал е най-късия числов интервал, който покрива неизвестната медиана с голяма точност (обикновено 95% или 99%).

Ще построим доверителни интервали за медианите на случайните компоненти X_i, i=1,…,7, на вектора X. Доверителните интервали са интервални оценки за съответните медиани на популациите (в случая медианите на количествените характеристики на сложните смесени изречения в романа „Антихрист" на Емилиян Станев).

В [2] са представени нови факти за езика на Емилиян Станев на базата на точкови оценки на числовите характеристики (квартилите на вероятностните разпределения) на количествените параметри на неговите изречения.

Нашата цел сега е да направим статистически изводи за езика на Емилиян Станев на базата на интервални оценки на количествените характеристики на неговите сложни смесени изречения. Нашите изводи ще се основават на една част от произведението – на случайна извадка от 200 сложни смесени изречения. Те не могат да бъдат 100% верни. Но когато не знаем *точно*, добре е да знаем *с вероятност*. Фиксираме максималната вероятност, с която можем да сгрешим в нашите изводи, да е малко число. Максималната вероятност, с която можем да си позволим да направим грешка за синтаксиса на Емилиян Станев е 0.01. Това означава, че нашите изводи ще бъдат верни с 99% сигурност.

Ще построим 99% доверителни интервали с команди на R и ще направим статистически изводи [3].

По този начин на базата на анализираните изречения и подходящи статистически методи ще направим статистически изводи за синтаксиса на цялото произведение „Антихрист" на Емилиян Станев.

4.1 Изследване на броя прости изречения в сложното (обхват) в произведението на Емилиян Станев

Броят на простите изречения в сложното е съхранен в променливата table$rank. С командата wilcox.test строим доверителния интервал.

```
> wilcox.test (table$rank, conf.level=0.99, conf.int=TRUE)
        Wilcoxon signed rank test with continuity correction
data:  table$rank
V = 20100, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 0
99 percent confidence interval:
 5.499966 6.000060
sample estimates:
(pseudo)median
       5.50005
```

Статистически извод: 99% доверителен интервал за обхвата е (5.5, 6.0).

С 99% сигурност можем да твърдим, че броят прости изречения в сложното смесено в това произведение е 5 или 6.

Извод за синтаксиса: Типичното за романа „Антихрист" на Емилиян Станев е, че броят прости изречения в сложното смесено (обхватът) е 5 или 6.

4.2 Изследване на броя главни изречения

По аналогичен начин правим изводи за броя главни изречения.

Статистически извод: 99% доверителен интервал за броя главни изречения е (2, 2.5). С 99% сигурност можем да твърдим, че броят главни изречения в сложното смесено в това произведение е 2 или 3.

Извод за синтаксиса: Типичното за романа „Антихрист" на Емилиян Станев е, че броят главни изречения в сложното смесено е 2 или 3.

4.3 Изследване на броя подчинени изречения

Статистически извод: 99% доверителен интервал за броя подчинени изречения е (3, 3.5). С 99% сигурност можем да твърдим, че броят подчинени изречения в сложното смесено в това произведение е 3 или 4.

Извод за синтаксиса: Типичното за романа „Антихрист" на Емилиян Станев е, че броят подчинени изречения в сложното смесено е 3 или 4.

4.4 Изследване на броя съподчинени изречения

Статистически извод: 99% доверителен интервал за броя съподчинени изречения е (2.0, 3.0). С 99% сигурност можем да твърдим, че броят съподчинени изречения в сложното смесено в това произведение е 2 или 3.

Извод за синтаксиса: Типичното за романа „Антихрист" на Емилиян Станев е, че броят съподчинени изречения в сложното смесено е 2 или 3.

4.5 Изследване на брой равнища (дълбочина) на изреченията

Статистически извод: 99% доверителен интервал за дълбочината е (2.9, 3.0). С 99% сигурност можем да твърдим, че дълбочината на сложното смесено изречение в това произведение е 2 или 3.

Извод за синтаксиса: Типичното за романа „Антихрист" на Емилиян Станев е, че дълбочината на сложното смесено изречение е 2 или 3.

4.6 Изследване на броя съчинителни връзки в изреченията

Статистически извод: 99% доверителен интервал за броя съчинителни връзки е (1.0, 1.5). С 99% сигурност можем да твърдим, че броят съчинителни връзки в сложното смесено изречение в романа „Антихрист" на Емилиян Станев е 1 или 2.

Извод за синтаксиса: Типичното за романа „Антихрист" на Емилиян Станев е, че броят съчинителни връзки между простите изречения в сложното смесено е 1 или 2.

4.7 Изследване на броя подчинителни връзки в изреченията

Статистически извод: 99% доверителен интервал за броя подчинителни връзки е (1.0, 1.5). С 99% сигурност можем да твърдим, че броят подчинителни връзки в сложното смесено изречение на това произведение е 1 или 2.

Извод за синтаксиса: Типичното за романа „Антихрист" на Емилиян Станев е, че броят подчинителни връзки между простите изречения в сложното смесено изречение е 1 или 2.

Заключение

Направихме статистически изводи за количествените характеристики обхват, дълбочина, брой главни изречения, брой подчинени изречения, брой съподчинени изречения, брой съчинителни и брой подчинителни връзки на сложното смесено изречение в романа „Антихрист" на Емилиян Станев.

Направените изводи за сложното смесено изречение надграждат съществуващите в българската синтактична традиция знания. Построявайки модел на количествените параметри на изреченската структура, настоящото изследване добавя нов подход за изследване на синтаксиса.

Благодарности

Проведените статистически изследвания за сложното смесено изречение в българския език са частично финансирани от проект ФП17-ФМИ-008 към Фонд "Научни изследвания" при Пловдивски университет "Паисий Хилендарски".

Литература

1. Бъркалова, П. Конституентна граматика и подчинени изречения. – В: Славистика IV. В чест на славистичен конгрес в Минск 2013. Пловдив: Университетско издателство „Паисий Хилендарски", с. 101-111.
2. Бъркалова П., В. Нончева, К. Колева. Граматическият формализъм и статистическият анализ в помощ на синтактичната теория и на синтактичната практика, МЕЖДУНАРОДНА ЮБИЛЕЙНА КОНФЕРЕНЦИЯ НА ИНСТИТУТА ЗА БЪЛГАРСКИ ЕЗИК „ПРОФ. ЛЮБОМИР АНДРЕЙЧИН" 15-16 май 2017г.
3. Димитров, Б., Н. Янев. Вероятности и статистика. София: Университетско издателство „Св. Климент Охридски", 1998.
4. Колева, К. Наблюдения върху синтаксиса на сложните смесени изречения в романа „Антихрист" от Емилиян Станев. Пловдив. Дипломна работа, 2016.
5. Станев, Ем. Антихрист. - Събрани съчинения в седем тома. София: Български писател, 1982, с. 175-330
6. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0, URL http://www.R-project.org/

STATISTICAL INFERENCE FOR QUANTITATIVE CHARACTERISTICS OF BULGARIAN COMPLEX-COMPOUND SENTENCE

Veska Noncheva, Petya Barkalova, Alexandar Ivanov, Vladimir Chavdarov

University of Plovdiv Paisii Hilendarski
24 Tzar Asen, 4000 Plovdiv

Abstract: *A new probabilistic model of sentence is presented. This model reflects the basic quantitative characteristics of the sentence. Statistical inference for Bulgarian complex-compound sentence in the novel Antichrist written by Emilian Stanev is made.*