S. Ahlborn

TOKENS AND TYPES DISTRIBUTION IN TITUS

Abstract. The present study examines the distribution of tokens and types in each text of the TITUS corpus depending on the source language. To determine this distribution a special program was developed which automatically calculates the number of tokens and types, integrated into TITUS and thus complementing its resource search engine. Further, the CMDI metadata set specific for TITUS resources is presented, which allows public access to the types and token distribution.

1. Titus Resource – general data and search engine

TITUS (Thesaurus Indogermanischer Text- und Sprachmaterialien) has been developed since 1987 thanks to the effort of Jost Gippert and can be reached online since 1994 via **http://titus.uni-frankfurt.de.** The TITUS database primarily includes texts of old Indo-European languages. It also contains texts of non-Indo-European languages (Caucasian languages). Mostly all of these resources are freely available. All texts are in HTML and in XML format (the latter ones will be online soon) and are encoded in Unicode/ UTF8. TITUS currently includes 660 texts in 55 languages, more than 30 m tokens[1].

The exact statistical information is necessary for proper scientific work with digital resources. Traditionally, corpora data are measured by the number of tokens and types. A token represents the concrete occurrence of the linguistic unit, and in a type, tokens associated with each other are bundled.

The TITUS Search Engine does not determine the number of tokens in the concrete text, but the number of quotations of the word. So it is possible to obtain by clicking a word in the text a summary of all quotations of the word in the current text. By this it will be searched both for the given word form and for its normalized version. For example, by clicking on *azй* азъ «i» in the Old Church Slavonic

[1] Estimation: April 2013.

text Marianus[2] 420 quotations will be obtained. In this example, two forms are also automatically calculated, referring to the lower and upper-case letters of the word. Another word click results in 1602 documents in the entire corpus of Old Church Slavonic texts. The total number of quotations is displayed at the end of the search results.

2. Peculiarities of TITUS texts

A closer look at the TITUS resources shows that a simple statement of the number of tokens can not be formulated. The reason for this is that the valuable online resources of TITUS often contain interlinear passages from parallel texts, editions and font variants.

For example, the text to the Gothic Bible[3] contains additional parallel passages in Latin and Greek: fig. 1.

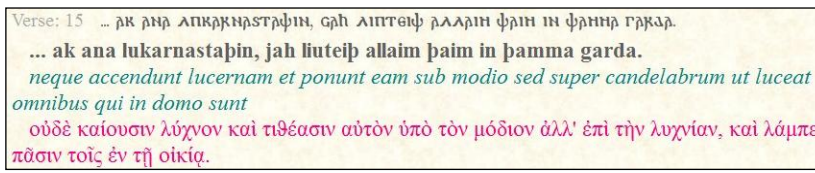

Fig. 1. Additional parallel passages in Latin and Greek in the Gothic Bible

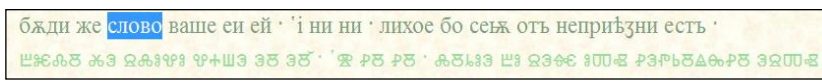

Fig. 2. Codex Marianus

Old Church Slavonic texts are represented in two ways: in the Glagolitic alphabet – original form of the text – and in Cyrillic one.

---

[2] **http://titus.uni-frankfurt.de/texte/etcs/slav/aksl/marianus/maria.htm**

[3] Biblia Gothica: **http://titus.uni-frankfurt.de/texte/etcs/germ/got/gotnt/gotntlex.htm**

However words in the text in Cyrillic are linked with the database: fig. 2[4].

Old Polish Texts[5] contain a simultaneous display of editions that have arisen at different times[6].

The Ossetian Nart epicis is represented in Latinica und in the advanced Cyrillic. Here both text variants are linked with the database: fig. 3[7].

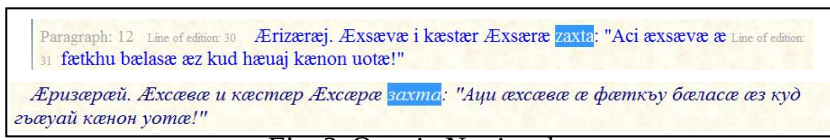

Fig. 3. Ossetic Nartic tales

The bilingual Russian – Low German text from the 17th Century[8] is challenging, for words of at least 9 different languages or language variations can be found (Old Russian in Latin transcription and in Cyrillics, Old Low German, Old Polish, Latin etc.): fig. 4.

The Old Prussian corpus consists of at least 21 different languages or language variants (Old Prussian, Old Lithuanian, Latin, Gothic, Old Low German, Old High German etc.)[9].

---

[4] Codex Marianus: **http://titus.uni-frankfurt.de/texte/etcs/slav/aksl/marianus/maria.htm**

[5] Kazania Świętokrzyskie**: http://titus.uni-frankfurt.de/texte/etcs/slav/apoln/kazania/kazan.htm**

[6] *P. Diels*. Die altpolnischen Predigten aus Heiligenkreuz, Berlin: Weidmannsche Buchhandlung, 1921; *J. Łoś / W. Semkowicz*. Kazania t.zw. Świętokrzyskie, Kraków: Polska Akademja Umiejętności, 1934.

[7] Corpus of Nartic tales**: http://titus.uni-frankfurt.de/texte/etcs/iran/niran/oss/ nart/nart.htm**

[8] Tönnies Fenne's Manual: **http://titus.uni-frankfurt.de/texte/etcs/slav/ aruss/ fenne/fenne.htm**

[9] Old Prussian Corpus: **http://titus.uni-frankfurt.de/texte/etcs/balt/apreuss/apreuss/apreu.htm**

The heterogeneous character of texts in TITUS promoted therefore the individual treatment in the calculation of tokens and types distribution.

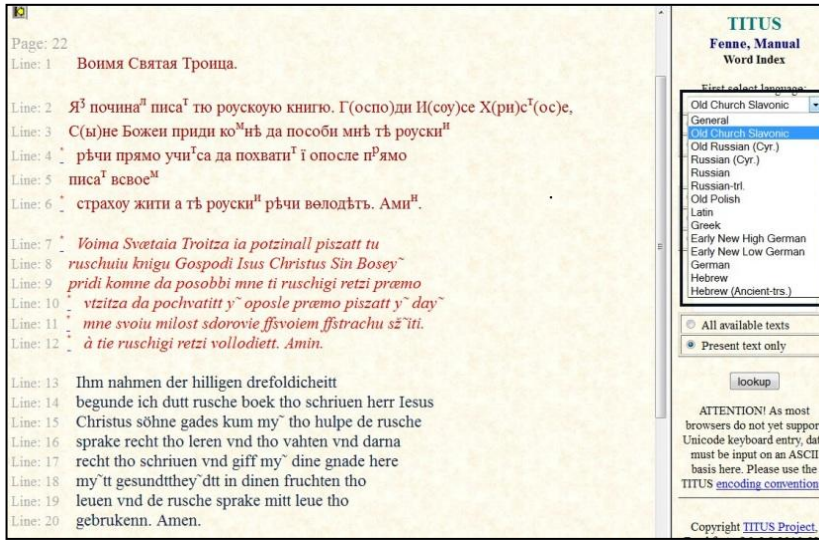

Fig. 4. Tönnies Fenne's Low German Manual of Spoken Russian

3. Tokens and Types calculation in the Titus Resource

To carry out the counting of tokens and types in TITUS resources several processing steps were necessary. Through numerous regular expressions in a Perl script the irrelevant information has been removed. A digitized source consists not only of a source language words, but contains various information which does not belong originally to the document: numbers, tags, punctuation marks, edition information etc. (e.g. <:>, <.b>, <Heidelberg>, <?ConvertCheck:)> etc.).

The further task was to assign automatically language names to correspondent word lists. The source file, extracted from the Wordcruncher database, contains data that are organized in various languages: «Language 1» may contain such words that do not belong

to the original text: «Heidelberg», «Bibel», «TITUS» etc. Starting from «Language 2» the list of words of the original text itself begins. In some sources, such as Gothic Bible in TITUS, «Language 3» relates to the Latin words and «Language 4» to the Greek ones. Depending on the language consistency of a text this list can be continued. Thus, the program calculates the distribution of tokens and types depending on the language a text consists of.

Here are examples of two TITUS resources, demonstrating the calculation results. So while calculating independently the language consistency of Old Testament Fragments of the Gothic bible in TITUS this resource consists of 1629 tokens. A calculation depending on the language shows, that this resource comprises 420 tokens | 240 types in Gothic, 572 tokens | 325 types in Latin and 627 tokens | 319 types in Greek. If the focus of investigation is on the linguistic material of Gothic, the relation between tokens and types of this particular language in comparison to the complete TITUS resource can be of great importance.

The complete Russian – Low German text from the 17th century Tönnies Fenne's Manual, representing the textbook of colloquial Russian, comprises the material of different languages or language variations. This language diversity can be explained by the fact that the manual contains in its lexical and grammatical part Russian in Cyrillic letters and in Latin transcription translated into Low German, phraseological examples in Russian in Latin transcription, grammatical terms in Latin and finally prayers and epistolary samples in Old Polish. This diversity is reflected by the distribution of tokens and types as follows: the complete resource consists of 55661 tokens and 19721 types, while according to the language diversity of this resource the following tokens and types distribution can be presented: 4838 tokens | 3613 types in Russian (Cyrillica), 21064 tokens | 7878 types in Russian (transliterated), 26201 tokens | 6104 types in Low German, 666 tokens | 509 types in Old Polish and 390 tokens | 258 types in Latin. According to this calculation it is possible to make a further conclusion. The language of the textbook of spoken Russian, which was created in order to support the bilingual conversation of

Russian and German traders of the 17th Century, consists mainly of Russian in Latin transcription and Low German.

The next step would be a more detailed study of the material. The TITUS database provides a list of words for further linguistic analysis. So, it is possible to get a list of words relating to the Latin words of the Tönnies Fenne's Manual, which were calculated with 390 tokens and 258 types, and find out where exactly the words of the Latin language were used in the text: fig. 5.

4. Metadata for Tokens and Types distribution

Finally, the distribution of tokens and types is issued in a specially developed XML file according to the CMDI (Component MetaData Infrastructure) metadata scheme. Metadata are data about data. They include information about properties of linguistic resources. Meanwhile, several initiatives have been established for this purpose: HTML, Dublin Core, IMDI, OLAC/DC, TEI, CMDI. CMDI characterizes the most important metadata scheme for language resources since it allows a flexible presentation of a resource and it is not bundled only to a bibliographic description[10]. Apart from using the ready-made components and profiles, it is also possible to create elements suitable for the individual resource. Being developed within the framework of CLARIN (Common Language Resources and Technology Infrastructure)[11], CMDI metadata can be used locally or they can be stored on the CMDI server. This allows the quick retrieval of data – both by users and by programs. For this reason the initiative CMDI was selected for the description of metadata for the TITUS resources. The new metadata set includes the newly created tags giving the information about the time of the original manuscript or facsimile and the distribution of tokens and types depending on the languages of the current text.

---

[10] CMDI: **http://www.clarin.eu/cmdi**

[11] CLARIN: **http://www.clarin.eu**

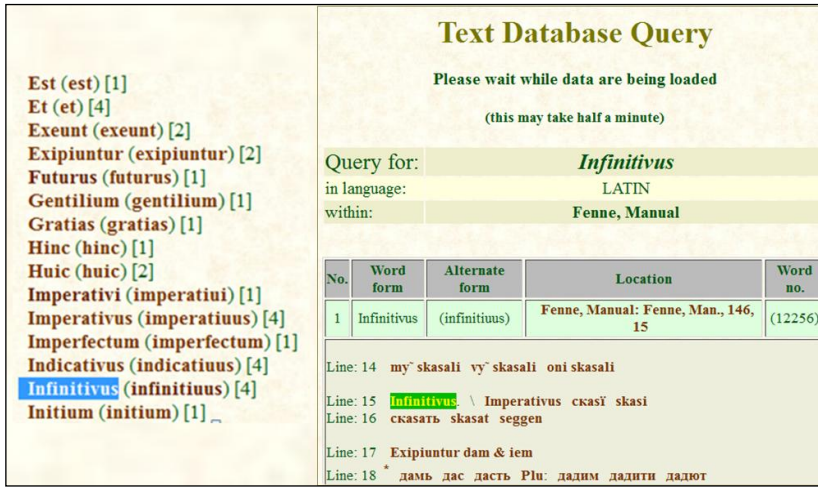

Fig. 5. Latin words in Tönnies Fenne's Manual:
further application of the tokens and types distribution

5. Conclusion

The study shows the results of the distribution of tokens and types in selected TITUS resources, depending on the languages of a given text. With the exact statistical information, the research within TITUS data can be now formulated more precisely. Conclusions can not only be made of the vocabulary size, but also of the richness and stylistic differentiation of each document or the type-token ratio of text collections can be compared.

The file containing the metadata information with the distribution of tokens and types can be output in various formats (e.g. XML, HTML). Since it is planned to continue to convert the TITUS resources into the XML format, the metadata have been created in the XML format. Metadata are already available online[12] and in the future will be found in TITUS next to the link to the text.

[12] **http://user.uni-frankfurt.de/~lana/Metadaten.html**