# TSCC: a New Tool to Create Lexically Saturated Text Subcorpora

**Zygmunt Vetulani, Marta Witkowska**
Adam Mickiewicz University in Poznań
Poland
vetulani@amu.edu.pl, martusiazielinska@gmail.com
**Umut Canbolat**
University of Kocaeli
Turkey
u.canbolat@yahoo.com

**Abstract**

In the paper we present a new tool to evaluate lexical saturation of text corpora, where lexical saturation refers to a state in which it is hard to find new lexemes outside the corpus. Estimation of the saturation degree for a given corpus contributes in a natural way to the corpus quality evaluation. We propose saturation tests as a stopping criterion for subcorpora creation. Although the first application of the TSCC tool is the evaluation of lexical coverage of corpora, it may be equally useful to study corpora representativeness for various phenomena, and – more generally – their usefulness for corpus-based research, both theoretical and practical (as e.g. studies of information impact). It may serve for cost evaluation of expensive engineering tasks in language competence modelling for AI purposes as well as in literary research. The system (TSCC) is highly language independent, i.e. it may be applied directly or easily adapted to any language in which the text units may be represented in alphabetic scripts. Its preliminary version (OCASSC) has been tested on a corpus of clients' opinions published by booking.com. The prototype will be freely distributed for beta testing.

**Keywords:** text corpora, corpora quality, lexical saturation tests, subcorpora creation, stopping criterion

## 1. Introduction

Although there is consensus about fundamental importance of linguistic data corpora (texts, recordings) for investigating natural languages according to the world-observation-based methodology of natural sciences, there is still a need of commonly accepted methods for text corpora evaluation. Initially, the size of the corpus was considered as a sufficient quality measure for corpora but quickly it has become clear that this is not an absolutely effective solution for corpora quality evaluation. The need of producing linguistic models for particular applications brought the attention of language engineers to specific linguistic phenomena. Consequently, corpora for language modelling are supposed to be *representative* for the phenomena in question[1]. As corpus collection is expensive (time, effort) and difficult (legal issues), quality evaluation of existing corpora is an important issue.

Representativeness of corpora was largely discussed in the wider context of corpora quality sometimes opposed to the concept of size as quality measure. In the frequently cited paper Douglas Biber (1993) presented, without any formal definition however, what it means to 'represent' a language. He considered various aspects of this concept taking into consideration language data stratification, sampling and – last but not least – size. Biber's work provides argumentation in favor of both qualitative and quantitative basis for corpus design (Kennedy, 1998).

In this study we follow the observation that "a huge corpus is not necessarily a corpus from which generalization can be made. A huge corpus does not necessarily 'represent' a language or a variety of a language any better than a smaller corpus" (Kennedy, 1998). This observation seems to be particularly adequate in a study of local lexicographical phenomena for which the question "how many tokens of a lexical item are necessary for descriptive adequacy" (ibid.) is justified.

## 2. Corpus Saturation

There exist various methods to estimate the representativeness of a sample of data for a given phenomenon. What they have in common is the evaluation of the chance of finding a new manifestation of the phenomenon outside the sample. In a representative sample all relevant examples should occur at least once. If the sample is (almost) representative then (almost) all newly observed manifestations will be identical to some already done. In particular, the size of the list of single manifestations of the phenomenon (list of hapaxes) will decrease or remain the same after each new observation. The decrease in the number of manifestations results from the increase in the corpus saturation with respect to the considered phenomenon.

## 3. Lexical Saturation

We will explore the concept of lexical saturation of a corpus (cf. e.g. Kittredge 1983, also Vetulani 1989). This concept appeared useful in research on the evaluation of the size of virtual vocabulary of sublanguages[2] (e.g. in the context of machine translation) and was used to study lexical saturation of corpora. Informally, we say that corpus is lexically saturated when "new lexemes appear only sporadically as a result of the extension of the corpus in a natural way" (Vetulani 1989).

In order to study the lexical saturation we consider the corpus to be a linearly ordered set of elements (words, symbols etc.). For its initial segments of the length N we observe the number of different words V. This function is increasing and the observation of the growth of V informs us about the degree of saturation of the corpus. The

[1] We consider a corpus as representative for a given language phenomenon, or a class of phenomena if it contains examples for all relevant aspects of this phenomenon.

[2] See e.g. (Kittredge 1983).

function may be represented graphically by a saturation graph. Observations of corpora confirm that V grows systematically with N but slower. The reason for this is that the observed vocabulary becomes more and more saturated, i.e. it is more and more difficult to introduce new words into discourse (Vetulani 1989).

For a sound[3] data gathering procedure it is crucial to have a good stopping condition, i.e. criterion to stop data collection. A good stopping condition will prevent against collecting data beyond necessity.

Let us consider the corpus as a linearly ordered partition into segments of equal size. Observation of the number of new words $\Delta V$ in the last segment informs us about the degree of lexical saturation of the corpus. It follows that a sufficiently small value[4] of the ratio $\Delta V/\Delta N$ is a good candidate for the stopping condition. Checking whether this stopping condition is satisfied is called *lexical saturation test*.

If we intend to compare saturation degree between corpora of different sizes it may be convenient to calculate the ratio $\Delta V/\Delta N$ for the last segment representing X% of the whole corpus (denoted $\Delta V/\Delta N(X\%)$ and called (*X% growth ratio*) for all corpora (and then to compare).

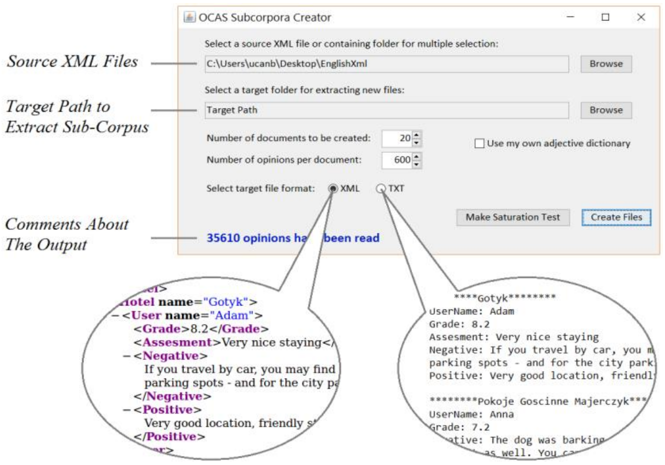
This method may be generalized in a natural way to evaluate representativeness of a corpus with regard to various phenomena. For example, in order to evaluate the minimum size of a balanced opinion corpus we performed experiments involving opinion adjectives (as adjectives are – for most of languages – the main lexical tool to support classification of opinions into negatives or positives).

Notice. Lexical saturation as stopping condition may be inadequate for corpus based studies of some global phenomena where statistical methods demanding huge amount of data or neural algorithms requesting large training sets are in standard usage. (See e.g. McEnry and Hardie (2012) or Peris et al. (2017)).

## 4. OCASSC

In 2017 we implemented the system OCASSC (Opinion Corpora Acquisition Software for Subcorpora Creation) initially designed to create corpora of opinion texts. In particular, it was used to randomly generate possibly small subcorpora of a large collection of texts that could be considered representative for studies of lexical instruments to express opinions. For the purpose of investigation of corpora representativeness for the given phenomena, OCASSC system was equipped with a functionality that enables execution of incremental saturation tests.

OCASSC requires two sets of input data. The first one is a corpus in an XML format. Such a corpus may be generated by the system OCAS (Vetulani et al. 2015). The second set of data is a predefined list of elements whose role is to restrict the search space. These are words included in the so called *reference list* that may (but do not need) appear in the investigated corpus and are formal indicators of the relevant phenomena (for example, opinion adjectives, i.e. adjectives that may be used to support an opinion, as illustrated below). The program retrieves the input data to create subcorpora of the length specified by the user and to perform saturation test for the predefined reference list.



Select a source XML file or containing folder for multiple selection: C:\Users\ucanb\Desktop\EnglishXml

Select a target folder for extracting new files: Target Path

Number of documents to be created: 20

Number of opinions per document: 600

□ Use my own adjective dictionary

Select target file format: ● XML ○ TXT

35610 opinions ha been read

```
 el>
 otel name="Gotyk">
 –<User name="Adam">
   <Grade>8.2</Grade>
   <Assesment>Very nice staying</
  –<Negative>
    If you travel by car, you may find
    parking spots - and for the city pa
   </Negative>
  –<Positive>
    Very good location, friendly s
   </Positive>
   ar>
```

```
    ****Gotyk********
UserName: Adam
Grade: 8.2
Assesment: Very nice staying
Negative: If you travel by car, you m
parking spots - and for the city park
Positive: Very good location, friendl

********Pokoje Goscinne Majerczyk***
UserName: Anna
Grade: 7.2
 ative: The dog was barking
   as well. You ca
```

Figure 1: The OCASSC main screen

### 4.1 Input Data Examples

OCASSC accepts data in XML format where tags are used to separate the meta-information from the opinion texts. The (simplified) example provided below presents the input compatible with the one used by the system OCAS (Opinion Corpora Acquisition Software) for collecting opinions (Vetulani et al. 2015).

<All>
<Review>
<HotelName>Gotyk House</HotelName>
<Positive>The welcome, the location and the wonderful helpfullness and charm of the staff were notable. Breakfast was simple and ample. We chose to go in the cold (-10C) and were perfectly warm.</Positive>
<Negative>Would have liked a kettle in the room but accept that fire considerations in such an old house prevented it.Q</Negative>
</Review>
<Review>
<HotelName>Hotel Kazimierz</HotelName>
<Positive>The location was just what we wanted, the room was clean and quiet and the staff were friendly and helpful</Positive>
<Negative>nothing</Negative>
</Review>
<Review>
<HotelName>Hotel Polski Pod Białym Orłem</HotelName>
<Positive>Room was large for European standards and the bed was so comfortable. Breakfast was good with a broad array of

[3] Data gathering procedure is considered *sound* with respect to a given objective, if it guarantees acquisition of all data necessary to reach this given objective.

[4] The value is to be fixed depending on the purpose of the corpus design and development.

foods, they also had good scrambled eggs and sausages. Would definitely stay here again.</Positive>
<Negative>Nothing, it was perfect for our stay in this beautiful city.</Negative>
</Review>
</All>[5]
(We have underscored opinion adjectives)

### 4.2 Output Data Examples

OCASSC was designed to generate subcorpora with the desired properties (saturation), so the basic format of the resulting subcorpora is the same as input. However, for more readability the system may output data in a text format.

```
*******Gotyk House*******
Positive: The welcome, the location and the wonderful helpfullness and charm of the staff were notable. Breakfast was simple and ample. we chose to go in the cold (-10C) and were perfectly warm.
Negative: Would have liked a kettle in the room but accept that fire considerations in such an old house prevented it.
*******Hotel Kazimierz*******
Positive:The location was just what we wanted, the room was clean and quiet and the staff were friendly and helpful
Negative: nothing
*******Hotel Polski Pod Białym Orłem*******
Positive: Room was large for European standards and the bed was so comfortable. Breakfast was good with a broad array of foods, they also had good scrambled eggs and sausages. Would definitely stay here again.
Negatative: Nothing, it was perfect for our stay in this beautiful city.
```

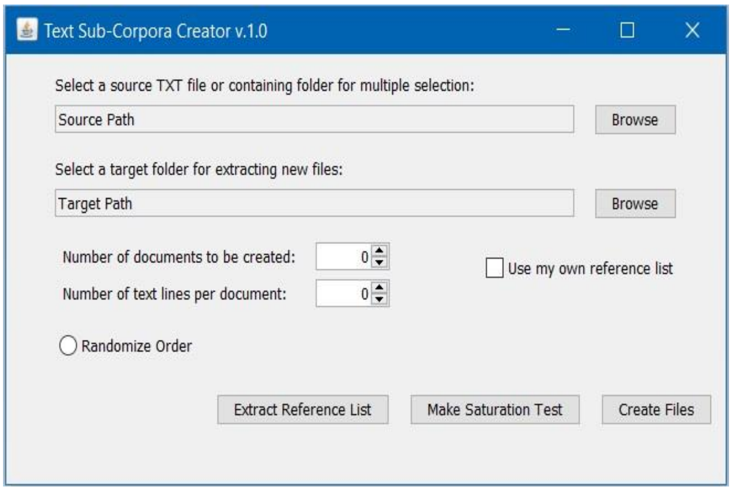
### 4.3 Experiment

Configuration of the OCASSC system requires a priori definition of the search space for the phenomenon of concern. In our experiment the search space was determined by the list of all adjectives that may be used to express opinion and which we consider interesting for our purposes. To create this list we useed a corpus of 2040 opinions in English (for hotels in the city of Poznań). We proceeded to manual annotation of all occurrences of adjectives used as opinion words. The next step was to create a frequency list of all annotated adjectives. This list contained 490 various adjectives (for 11854 occurrences). 312 adjectives that occur more than once in the corpus were used in the experiment as reference lists (we discarded hapax legomena in order to limit the number of atypical opinion words in the reference list). Then we applied OCASSC to a corpus of opinions (in English) about hotels in Poland containing over 850.000 of text words (34.800 opinions containing 28.371 different words) in order to extract subcorpora of 2040 opinions. We applied the *10% growth ratio* (with respect to opinions) to evaluate the degree of saturation for these subcorpora. In each of the observed cases the ratio varied between 0.01 and 0.03.

On the other hand, application of OCASSC to the corpus of 34.800 opinions and to the set of *all words* as reference list, shows that the corpus is far from being lexically saturated as the *10% growth ratio* for the reference list containing all (general) vocabulary is relatively very high (0.453). Consequently, a significant growth of the observed vocabulary should be expected when proceeding to its extension.

## 5. TSCC – Text SubCorpora Creator

The purpose of the software is to create smaller corpora from a large text corpus. Reusing the OCASSC system architecture, we have designed and implemented (2017, summer) a system operating in open (unformatted) text as a tool for extracting subcorpora of any desired size from a large corpus. In situations where the representativeness of the corpus is closely related to its lexical completeness (which is the case of sublanguages determined by restricted application domains, such as e.g. the sublanguage of metereological reports[6]) evaluation of the degree of lexical saturation using TSCC may help to fix the stopping criterion for creation subcorpora with desired properties.



Select a source TXT file or containing folder for multiple selection: Source Path

Select a target folder for extracting new files: Target Path

Number of documents to be created: 0

Number of text lines per document: 0

□ Use my own reference list

○ Randomize Order

Figure 2: The TSCC main screen

### 5.1 TSCC v.1.0 Functionalities

The system processes the input corpus in the form of a text(.txt file). The "Source Path" browse button will permit the user to provide location and open the input text. The input text may have a form of just one or several input files contained in a folder. Selection of the source file or folder will result in evaluation of the whole text (corpus) and the total number of lines will be displayed. The next step is to select the target folder like we did for selecting the source

[5] Spelling and syntax are original.

[6] The studies of the sublanguage of metereological reports were the object of special interest in the classical linguistic and AI literature ((Muller (1975), Kittredge (1983)).

path. The output files will be put in a folder created by the system.

After the input/output operations, the remaining processing parameters must be declared. The corpus text will be considered as composed of text lines grouped into “documents” containing fixed number of lines. This number, as well as the number of documents are to be specified by the user. These two numbers will determine the length of the subcorpus extracted and processed by TSCC. Creation of documents may be done line by line (default solution) or in a random way with respect to the whole input corpus.

“Reference list” plays the same role as in OCASSC: it defines the set of elements (words or special tags) to be taken into account to define the saturation function (and therefore determine the processing search space). The user is supposed to use his/her own reference list (“Use my own reference list” checkbox).

Finally, the user is supposed to declare one of the following three functionalities: to create a subcorpus of the specified size (button “Create files”), to calculate data for the saturation test (button “Make Saturation Test”), or to transform the input corpus into a reference list composed of all word forms of the corpus[7] (button “Extract Reference list”). The results are being saved in the output folder.

### 5.2 Experiment

TSCC generates data necessary to draw the saturation graphs for text corpora (function “Make Saturation Test”). Figure 3 presents the graph for a corpus composed of 8920 lines randomly extracted from a larger corpus of hotel opinions (given by Booking.com clients). The corpus appears not large enough to be considered as lexically saturated (for adjectives) because the increase of numbers of adjectives is quasi linear after the first 4000 lines. Still it is not very high, as the local increase speed expressed by the 10% ratio equals 12 words per 1000 lines). (The 10% ratio with respect to the number of text words equals 0.0012).[8]

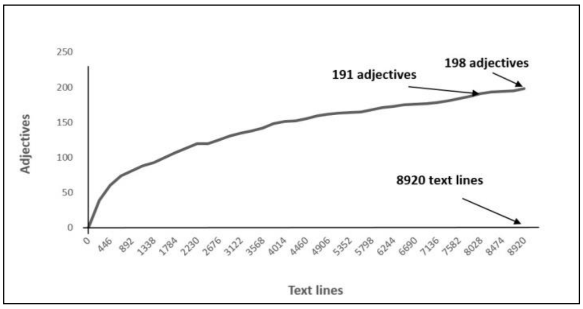


Figure 3: Saturation graph for all adjectives observed in the corpus of 8920 text lines (hotel opinions)

## 6. Possible Application in the Information Impact Studies

Measuring information impact is an important practical issue in many contexts: political, social (public security), military, etc. Impact is measurable as far as information is registered and stored in a systematic way, e.g. in form of text corpora. As it is a rule in empirical studies, appropriate sampling[9] is crucial. TSCC, or equivalent tools, may be useful in this venture but requires a careful selection of reference lists of words as information filters. Examples will be presented and discussed at the workshop. Below we limit ourselves to the main ideas only.

Impact of an event may be estimated on the ground of the registered information in the form of text. We will assume that information we are interested in is represented in a text corpus (called basic corpus), e.g. in a corpus of press news, and that the event description may be identified by a list (reference list) of terms (or concepts) (possibly including proper names, acronyms, dates etc.). The reference list must be individually defined for the events in question.

If an event may be identified (on the ground of the reference list) through a search in small, random selected samples (subcorpora) extracted from the basic corpus, (i.e. it is easy to find in the basic corpus) then we will be entitled to conclude that its impact is big. In practice, generation of the samples (subcorpora) may be done using the TSCC system for a predefined stopping criterion in order to guarantee the appropriate saturation of the generated samples by the elements of the reference list.

In a similar way we may evaluate what is the part of a well-defined subject area (with respect to other subject areas) in the literary output of an author. In that case the whole literary production of the author (or its representative fragment) constitutes the basic corpus, and lexical formal subject indicators form the reference list. It is in our imminent plans to apply this method to analyze the literary works of Polish writer and poet Julia Hartwig (https://en.wikipedia.org/wiki/Julia_Hartwig) and to present the results at the CIDTD Workshop at LREC 2018.

## 7. Further Research

We intend to further develop the TSCC system from the point of view of literary research. In particular its utility will be beta tested in the research on vocabulary structure of particular authors and particular literary works. We hope to prove its utility for stylometry. The beta prototype of the TSCC tool will be released for LREC 2018 and distributed under an open license.

[7] This may be useful in order to evaluate the degree of lexical saturation of the corpus with respect to the whole lexicon, e.g. using the X% ratio method described above.

[8] The Reader should however be warned that the saturation parameters in the experiments of sections 4.3 and 5.2 ought not to be confronted, as the reference lists (adjectives in both cases) are different. In particular, the used reference lists are different and the Princeton WordNet list of adjectives used in 5.2 is very poor in opinion-supporting adjectives.

[9] By *sampling* we mean here selection of subcorpora of appropriate size (small).

## 8. Acknowledgement

We intend to thank Jolanta Bachan, Katarzyna Klessa, Gerard Ligozat, Patrick Paroubek and Suleyman Menken for their precious assistance and help.

## 9. References

Biber, D. (1993). Representativeness in corpus design, *Literary and Linguistic Computing,* Vol. 8, Nb. 4, pp. 243–257.

Kennedy, G. (1998). *An Introduction to Corpus. Linguistics*. Longman: London and New York.

Kittredge, R. (1983). Semantic processing of texts in restricted sublanguage. *Computers & Mathematics with Applications,* Vol. 9, Issue 1, pp. 45–58.

McEnry, T., Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice.* Cambridge University Press.

Muller, Ch. (1975). Peut-on estimer l'étendue d'un lexique? *Cahiers de Lexicologie,* Nb. 27, 1975–II, pp. 3–29.

Peris, Á., Chinea-Ríos, M., Casacuberta, F. (2017). Neural Networks Classifier for Data Selection in Statistical Machine Translation. *The Prague Bulletin of Mathematical Linguistics*, Vol. 108, June 2017, pp. 283–294.

Vetulani, Z. (1989). *Linguistic problems in the theory of man-machine communication in natural language. A study of consultative question-answering dialogues. Empirical approach*. Brockmeyer: Bochum.

Vetulani, Z., Witkowska, M. and Menken, S. (2015). Corpus Based Studies on Language Expression of Opinions. In: Z. Vetulani and J. Mariani, editors, *Proceedings, 7th Language & Technology Conference, Human Language Technologies as a Challenge for Computer Science and Linguistics, Poznań, 2015*. Fund. UAM: Poznań, pp. 365–369.