Chapter 2

Model of the Empirical Distribution Law for Syntactic and Link Words in “Perfect” Texts

Pavel Makagonov
Russian Presidential Academy of National Economy and Public Administration, Russia

ABSTRACT

The measure of perfection of the contents and semantic value of an integrated text is connected with the indicators of perfection in the distribution of content words. This criterion is the coordination of their “frequency-rank” distribution with the Zipf or Zipf-Mandelbrot law. In this chapter the hypothesis verified is that a perfect system should have not only perfect distribution of its elements - objects, but also perfect connections between them. A model is suggested in which the degree of the text perfection from the point of view of the quality of connections between significative words is determined by the quality of distribution of syntactic and link words in the “rank - frequency” representation. As a simplified criterion the ratio of the significant and syntactic words used in the analyzed text and the degree of the closeness of this ratio to the “golden section” is considered.

INTRODUCTION

In this study we shall consider an integrated text, a complete document or a work of art as a system the degree of perfection of which is estimated by the end user (reader, consumer of the information contained in the document). Treating a text as a semantic structure raises the problem of the text content segmentation, with various units suggested to isolate the quantum of the text information. As shown in a number of studies, there is a hierarchy of semantic addend found in the text, the latter “exists as a hierarchical unity breaking down into more and more fractional components” (Babenko and Kazarin, 2004). In this aspect, it is possible to consider not only the semantic importance of significant words, but also the influence of syntactic words on the meaning and content of the text. A model is suggested.
in which the degree of the text perfection from the point of view of the quality of connections between significative words is determined by the quality of distribution of syntactic and link words in the “rank - frequency” representation. The hypothesis verified is that a perfect integrated text should have not only perfect frequencies distribution of significative words, but also perfect connections between them.

There is a common opinion that in mathematical text models the semantically significant part of the text – content words have an unlimited tendency to enrichment and development, and the set of syntactic words is conservative. That is why it is important as well to study ratio of content significative) and syntactic words in “perfect” texts.

GOLDEN SECTION IN THE RATIO OF CONTENT AND SYNTACTIC WORDS IN “PERFECT” TEXTS

Let’s analyze now several integrated complete texts recognized “ideal” in the world literature. These are internationally acclaimed written masterpieces, oral folklore products written down, for example, fairy tales.

Basically, here there can also be collections of scientific texts on narrow subjects, exhaustively reflecting the subject of study.

Thirteen most well-known and popular texts which can be regarded as “systems close to perfection” have been analyzed:

- Homestead Act of 1862 (U.S. Congress, 1862).
- Кэрролл, Л. Приключения Алисы в стране чудес (Russian text Carroll, L. Alice’s adventures in wonderland, in translation N. M. Demurova, 1991).
- Carroll, L. Alicia en el país de las maravillas (Carroll, 2003).
- The Valiant Little Tailor – a fairytale (Grimm & Grimm, n.d.).
- Храбрый портняжка – Russian version of The Valiant Little Tailor Grimms’ Fairy Tales (Grimm & Grimm, n.d.).
- El sastrecillo valiente (Siete de un golpe) – a short story (Grimm & Grimm, n.d.).
- Das tapfere Schneiderlein – a fairytale (Grimm & Grimm, n.d.).
- The Three Little Pigs (Steel, n.d.).
- Сказка про трёх поросят (The Three Little Pigs Russian version) (Talebook.com, n.d.).
- Вересковый мёд (Heather Ale Russian – a ballad by R. L. Stevenson (Marshak, n.d.).
- Pushkin, A. C. A Tale about a Fisherman and a Fish (Russian text), (Pushkin, 1960).

The collection of texts includes four fairy tales and one ballad in English, Russian, German and Spanish, and additionally one old globally world known legal act. This minimum multilingual corpus was specially formed for detailed comparison and statistical analysis of the text of similar content and, consequently, similar in the semantic content of the texts, the corpus so formed to be used to check the hypotheses concerning the model of distribution of syntactic and link words and their role in the enhancement of the text as a semantic system.
19 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the product's webpage: www.igi-global.com/chapter/model-of-the-empirical-distribution-law-for-syntactic-and-link-words-in-perfect-texts/133874?camid=4v1


Related Content

How to Use Manual Labelers in the Evaluation of Lip Analysis Systems?
www.igi-global.com/chapter/use-manual-labelers-evaluation-lip/31070?camid=4v1a

Multiword Expressions in NLP: General Survey and a Special Case of Verb-Noun Constructions
www.igi-global.com/chapter/multiword-expressions-nlp/70061?camid=4v1a

Linguistic Certainty in Managerial Announcements
www.igi-global.com/chapter/linguistic-certainty-in-managerial-announcements/108719?camid=4v1a

Cindy K. Chung and James W. Pennebaker (2012). Applied Natural Language Processing: Identification, Investigation and Resolution (pp. 206-229).
www.igi-global.com/chapter/linguistic-inquiry-word-count-liwc/61050?camid=4v1a