Algorithmic Interpretation
A model of the reading process

em.o.Univ.Prof. Dr. Andrew U. Frank, Geoinformation, TU Wien

blog

1 Introduction

An attempt to define "literature" and "literary texts" suggestes "a literary text is a text which is written to be read"; the important aspect of this definition is what it does ot state: a literary text is not trying to convey some facts about the real world — it is fiction — and not to instruct, order, regulate or having some other effect in the real world. It just written to be read, maybe to amuse, entertain or otherwise change our emotions - but this is not the decisive property. Decisive is that is is written to be read and nothing else.

This definition of literary text seems to provide a convenient starting point for what is often termed "digital literature studies" (or, german, "digitale Literaturwissenschaft"), summarily including all efforts to apply computer analysis to literary texts. Pioneered perhaps by Moretti [1999] with three examples what a computer aided literary analysis could yield. The field has taken off since and a substantial number of literary studies using information processing in the analysis in one form or other, demonstrating many different tpes of analysis which can be aided by automated processing of the texts, falling into different types of analysis (e.g. stylometrie, ...).

In this contribution, I want to take a different approach to digital literature studies, focussing on the process of reading a literary text. The result of the reading process is the "interpretation" of the text by this reader; the interpretation is what this reader takes from the text, what the text tells him and what he may retain for future reference, what may him influence in future decision or actions. It is obvious, that different readers will arrive at different interpretations; the interpretation depending on the knowledge a reader brings to bear on the text during the reading process.

I suggest, that the reading process of a person reading a literary text can be modelled by a set of automated processes. The model clarifies the different steps a human reader performs to read a text. Breaking down the reading process in steps and – this is perhaps the major contribution – allows to identify the types of knowledge that is used in the natural reading process. It may then become possible

to connect an interpretation with the knowledge a reader must bring to the reading process, differentiating different interpretation with the previous knowledge a reader must have to arrive at this interpretation.

In the following, the different steps in reading and thus different types of analysis are separated. For each step the knowledge a reader must have in order to perform is detailed. In order to procude reproducable interpretations, the body of knowledge needed this interpretation must be described and i will indicate some ideas how the used knowledge can be characterized.

2 The reading process: Start at the text

A reader starts with a text which is in some form encoded - in many languages in characters or glyphs; nowadays, texts are often produced, reproduced and distributed in a computer encoded form. This means, that an algorithm analysing the text and the human reader start with the same material. The current standard UTF-8 is flexible and allows, in principle, the encoding of glyphs of any language; it allows for 1,112,064 valid code points, which each can stand for a glyph (i.e. a graphical represenation of a character). The initial text is a sequence of code points (in the sense defined by UTF8), for which I will use the familiar term “character”.

Excluded from consideration are pictures, drawing and similar parts in a literary work, which are not expressed as UTF-8 code points. Their integration into an analysis remains for later.

3 Reading language

A reader makes an intuitive decision on the language a text is written in and apply his understanding of the selected language to the text. The selection of language influences several of the following steps:

3.1 Tokenization

The text as a sequence of characters must be broken into meaning bearing words (or the equivalend of meaning carrying elements in other languages). The rules for breaking a text into words are language specific and more complex than just breaking at spaces. Think of composed words, words broken by end-of-line hyphens etc. This is achieved by language specific algorithms, which convert a text (i.e. a sequence of characters) into a sequence of tokens, each standing for a word in the language selected for the analysis.

Because tokenization rules for different languaes are different, the number of words in a text changes depending on the language specific tokenization selected; this makes stylometric analysis, which often use wordcounts, different to compare across languages.

3.2 Analyse form

The initial text is laid out on pages; the layout is encoded in the text with special "characters" standing for lline-breaks, page breaks and similar which are set by the author and are part of the text layout, not just accidentl line-breaks to fit a page width or length. The analysis of the form is concentrating on the distribution of tokens between the layout makrs:

- number of tokens between intentional line marks
- number of tokens and lines between intentional page breaks
- other layout marks

The structure of the text in pieces starting on new pages (maybe called chapters or poems) and the structure of text arranged in lines is preserved for further analysis. Computing median values for the number of tokens per lline and lines per page may help to separate traditional categories of literary texts into poetry, play, novel.

3.3 Knowledge used

The knowledge required is the set of rules to break a text into tokens, which is typically incorporated in the tokenizer software, e.g. the PTB tokenizer from Stanford University [1] suitable for english texts. Similar tokenizers for other languages have been developped by other computer linguistics groups.

4 Rhyme and alliteration analysis

A text, especially texts with intentional linebreaks introduced by the author may contain specific phonetic structure. To detect all structures in a text based on its spoken form requires a translation of the written text to is phonetic form.

4.1 Rhymes and alliterations

The sounds of the text at the beginning or the end of the line my sound similar. A reader with knowleadge of the prononciation of words in the language will identify these similarities. An automatic detection is possible, with observing how well the two pieces of text correspond (e.g. rhymes which appear forced, like ...).

4.2 Rhyming schemes

Some common verse forms are bound to a specific pattern of rhymes, e.g. sonnets rhyme ...

[1] http://nlp.stanford.edu/software/tokenizer.html

4.3 Prosody

The automatic idenfication of a meter and rhythm from a phonetic text means to apply a verse structure to a text and identify the amount of correspondence.

4.4 Knowledge required

For a reader to detect rhymes knowledge of the pronounciation rules of the language are necessary, an algorithmic interpretation can use a phonetic dictionary, e.g. the one compiled for American english by Carnegie-Mellon University [2].

To identify verse structures, both for rhyme structure as for verse, the reader has to know the verses structures most likely occuring and then test for it. The same can be done algorithmically, with some allowance for inexact matches, in rhymes, rhyming schemata and prosodiy.

5 Language analysis

The reader must connect the identified tokens with his personal understanding of the meaning of the words. Semantics of words are defined by their connection to real world

5.1 Narrative structure

6 Coda

tesxt

References

Franco Moretti. *Atlas of the European novel, 1800-1900*. Verso, 1999.

[2] http://www.speech.cs.cmu.edu/cgi-bin/cmudict