Structure Annotation in the Polish Corpus of Suicide Notes

Michał Marcińczuk[1], Monika Zaśko-Zielińska[2], and Maciej Piasecki[1]

[1] Institute of Informatics, Wrocław University of Technology,
Wybrzeże Wyspiańskiego 27, Wrocław, Poland
{michal.marcinczuk, maciej.piasecki}@pwr.wroc.pl
[2] Institute of Polish Philology, University of Wroclaw,
pl. Nankiera 15, Wrocław, Poland,
monik@uni.wroc.pl

Abstract. Polish Corpus of Suicide Notes (henceforth PCSN) is constructed to meet the needs of forensic linguistics. Suicide notes are messages created in borderline situation, shortly before death. Hence the annotation schema requires a complex description of a document structure, the textual content, as well as its linguistic properties. TEI was selected as the basis for the document encoding schema. TEI adaptation and extension with respect to such aspects of encoding as: a letter structure, various layers of changes and omissions, error correction, and extra-linguistic elements etc., are discussed with examples.

Keywords: forensic linguistics, suicide note, structure annotation

1 Introduction

Forensic linguistics is a branch of applied linguistics and one of its important applications is preparing linguistic evidence for the court. An opinion of a linguist who appears in the court as an expert witness must be based on the objective linguistic data not on intuition. General corpora are used as a referential material for comparison [1], however construction of specialized corpora for particular types of forensic texts is the key issue for the forensic research [11]. Forensic corpora are typically small, include short texts and are used for e.g., authorship attribution and text authenticity analysis. Comparison within the same type of text is the key tool in the description of idiolect focused on separating text type features from the personal linguistic features of the speaker.

Our main objective is to develop an annotated corpus of suicide notes (the last messages written by a person before committing a suicide; the message can be in a form of letter, note, SMS or other), which is intended to become a reliable data source for linguistic analysis of new individual suicide notes. The analysis can be later applied in the court.

Application of computational methods to suicide notes started with the work of Shneidman [2], who investigated differences between genuine and simulated notes. Two sets of suicide notes (33 genuine and 33 simulated) were compared

by using a system of categories which encompassed a set of predefined tags referring to roles, objects, emotional states, actions, institutions, statuses, qualities, symbolic referents [3]. These two sets of suicide notes were also used in [4]. Recently, Pestian et al. [5] applied machine learning methods to the recognition of false suicide notes. Their results are very positive, but their claimed superiority in comparison to the accuracy of humans causes doubts concerning the used evaluation method. "Vienna Corpus of Suicide Notes" was aimed at building psycholinguistics descriptions. It consists of suicide notes collected from years 2002–2005 and was used for comparing suicide note-writers with suicide non-note-writers. The comparison was based on eight variables: age, gender, marital status, occupation, psychiatric care, suicide motive and suicide method [6]. Additionally, analysis of suicide notes by forensic linguistic services is also a part of the general forensic document examination (e.g. ALIAS: software for forensic linguistic analysis[3]) performed with the help of the police corpora for crime investigation (e.g. suicide notes corpus from British Transport Police [7]).

A suicide note is on average a rather short piece of writing which is thematically and stylistically varied. Instead of performing a massive statistical analysis we have to strive for every piece of information characterising the text. Besides pure linguistic features, e.g. lexical or syntactical, pragmatic or even extra-linguistic features, e.g. the text structure and layout, can be also a valuable information source. Thus the annotation of a suicide note should encompass both layers: linguistic and structural. The latter is aimed specifically for forensic text analysis and support for tasks like suicide prediction. Structural information is mostly neglected in existing annotated corpora of suicide notes. Linguistic variables, like parts of speech and lexical frequency, as well as structure variables commonly used in quantitative text analysis, like the number of paragraphs and sentences, the length of sentences, etc., must be covered by the corpus annotation. However, we want to broaden this set. Only a subset of the functionality of the existing software for the forensic handwriting examination can be adapted to our task. For instance Wanda Workbench software supports annotation of content, material, script and writer but without structure annotation. The only elements of the structure that can be described in Wanda are characters and their selected measurable features. Text segmentation is still a significant problem in the forensic document pre-processing with OCR system [8].

The proposed suicide note structure annotation was inspired by three factors: handwritten form of the text, current forensic practice and requirements of the given text genre. We considered two types of text: suicide notes and Polish personal letters which are the most familiar type of an informally written text.

2 Choosing Text-encoding Standard

Text-encoding format should facilitate the intended use not determine it. Thus, before deciding about the particular encoding standard to be selected, we have identified several aspects that should be covered by the annotation:

[3] http://www.aliastechnology.com

1. Text structure and layout:
   - formal letter structure: opening, body, closer,
   - physical text division into text blocks and lines, e.g., paragraphs, marginalia, page and line breaking etc.,
   - text block layout, text alignment, indention, relative position,
   - text formatting, e.g., bold, italic, underline, etc.,
   - text omissions, deletion and insertion,
2. Correction
   - text correction introduced by authors and editors,
3. Linguistic information
   - segmentation into tokens and language expressions of various complexity,
   - semantic and pragmatic classification of text elements, e.g., salutation inside text, signature, envelope date expressed in different ways,
   - proper names,
4. Meta-data:
   - information about author,
   - physical description (paper format, type, etc.), linguistic description (type of text, e.g., letter, part of web log, statement).

In the above classification the most important seems to be the distinction between the description of the structure marked visually and linguistic properties of language expressions, that are independent from the former, e.g. salutation can occur as embedded inside a paragraph, not only in a separate text block.

We considered several standards for text representation, like TEI P5[4], XCES[5], KAF (*Kyoto Annotation Format*)[6] [9] and ISO TC 37/SC 4[7], as well as several formats developed for specific projects, like SCOTS or CEEC. Finally, TEI P5 was selected as it conforms to most our requirements and provides guidelines for manuscript description. TEI P5 also allows for the description of both the word-level and the medium of the document. It facilitates annotation of text segmentation, additions, ornaments, figures, underlining, crossing out, etc. TEI provides generic encoding guidelines for personal letters (among other genres) that can be further specified, e.g. with respect to cultural characteristics. This path was followed, e.g., in DALF[8], *Repertorium* project[9], *Vincent van Gogh - The Letters*[10], DBNL[11] and CARDS[12]. TEI allows to create the structure annotation which takes into account handwriting features, need for text reconstruction and writer identification. We follow this approach and use TEI as a basis for our corpus encoding. As not all of our requirements are met, e.g., handling hyphenated words, a further extension will be proposed.

[4] http://www.tei-c.org/Guidelines/P5/
[5] http://www.xces.org/
[6] http://xmlgroup.iit.cnr.it/kyoto/?option=com_content&view=article&id=141
[7] http://www.tc37sc4.org/
[8] http://www.kantl.be/ctb/project/dalf/
[9] http://clover.slavic.pitt.edu/repertorium/
[10] http://vangoghletters.org/vg/
[11] http://www.dbnl.org/
[12] http://alfclul.clul.ul.pt/cards-fly/index.php?page=mainen

3 Annotation Scheme

Annotation scheme is organised along several layers: physical layout (see Sec. 3), segmentation and morphological description[13], meta-data and semantic annotation.

Our starting point for encoding the physical and logical structure of the letter was DALF adaptation of TEI [10]. The letter body is divided into three parts: letter opener (`<opener>`), letter content (paragraph sequence `<p>`), letter closer (`<closer>`). Additionally, suicide notes are sometimes found in envelopes. More frequently a note includes the first page with information about recipient and the way of the suicide note delivery. Both are encoded by <envelope> tag: the envelop and the first page. In general the main parts are block elements and enclose complete lines of text. In some cases the opener and closer have a non-standard form, i.e. they overlap with a content paragraph, e.g., the first paragraph starts in the same line as the opener. To encode this `rend="inline"` attribute was used for opener and closer tags.

`<opener>` block can include three block elements: `<dateline>` – a date line occurring on the top of the letter, `<head>` – a title line and `<p>` – any other piece of text included in the opener. The last one can be repeated. `<closer>` can include three types of elements: `<dateline>` – in some letters the dateline appears not at the beginning but the end of the letter (but only once), `<p>` – a text passage (one and more) and postscript (`<ps>`). `<ps>` can encompass one `<p>` block element with two possible interpretations: `<p type="meta">` (*meta-paragraph*) to encode how the postscript section is introduced and `<p>` for a text block included. In some notes the meta-paragraph appears in the same line as the first postscript. Those cases are encoded by `rend="inline"` in the meta-paragraph. A postscript is both: a text block marked by the writer as PS, as well as, a text occurring after the signature. A postscript can be also introduced by a description, e.g. *I am adding*, *I'm also adding*, etc.

The line breaking `<lb/>` can appear in paragraphs (`<p>`) but also in date lines (`<dateline>`), when a date or a city name does not fit in one line, and in the title lines (`<head>`). Other elements consists of blocks, like `<opener>`. Page breaking inside `<p>` is expressed with `<pb/>` tag. It closes any open block element, i.e. divides one continuous text paragraph into two `<p>`. Specific visual separation of paragraphs (e.g. *a drawn line*) is described by <ornament/> tag, whose type attribute expresses the shape, e.g., *line*, *space*, *wave*, etc. (open list).

The horizontal alignment of text in block elements: `<p>`, `<head>`, `<dateline>` is stored in `rend` attribute with the following values: *left*, *centre*, *right*, *indent*, *margin-left*, *step-left*, *step-indent* and *step-center*. The `step-*` value describes the case in which the following line have bigger indention then the previous one. This is a characteristic feature of the Polish handwritten personal letters. The layout description of a block includes its positions as no location can be assumed as the default one.

[13] Morphological description will be included in the final version of the corpus.

Finally, pieces of text added in different places on the page (marginalia, doodles, etc.) are called *additions*. As the suicide notes layout has an atypical form (because of the context of writing e.g., lack of paper and emotional situation) we have to use variety means for its description: additions and paragraphs with the specified position which were written after the main text had been closed.

Not all parts of handwritten text can be read with enough confidence or are text in fact (e.g., drawings, signatures, etc.). All illegible fragments are annotated by `<gap/>` tag with four sub-types: *illegible* – a fragment impossible to read, a part is missing for some reason, *prosecutor* – a fragment obliterated by prosecutor (due to *anonymisation*) and *signature* – author's signature. If a text can be read but with some uncertainty, it is marked by `<unclear>` tag with a specified level of certainty (*low*, *medium* or *high*).

Any symbols or drawings that are not a sequence of characters are represented with `<figure/>` tag with type attribute describing the shape, e.g. *arrow*, *cross*, *emotikon*, *heart*, *other*, etc. (an open list).

For text replacement, deletion, addition etc. we use a combination of `<del>` and `<add>` tags. In contrast to `<gap>` the `<del>` tag is used when a piece of text is strikeout in some way but still readable. `<add>` (in contrast to `<additions>`) is used for a text inserted, e.g. between words or instead of strikeout word).

We distinguished two types of corrections made by: the author during writing (self-correction) or the editor during transcription. Editor corrections are important for automatic text processing. They are annotated by `<corr>` tag which encodes also the type of misspelling with several types predefined in the annotation guidelines. This description facilitates evaluation of writer's spelling competence.

Both correction types are distinguished by `resp` attribute with two values: *author* and *editor*. In the case of editorial corrections the original text is kept in `sic` attribute and the corrected text is put inside the tags.

The text formatting is described by `<hi>` tag with rend attribute encoding the formatting type, e.g. *bold*, *italic*, *underline*, etc.

For the needs of automated text analysis the correct text flow must be encoded, i.e. which fragments form a continues text. Two problems appeared: paragraphs divided between pages and word hyphenation. Concerning the former, we assume that page break breaks block elements. Thus, a continuous paragraph according to the authors intention is divided into two `<p>` tags, a kind of 'technical' paragraphs. Both are joined with the help of TEI aggregation mechanism and the `prev` and `next` attributes that points to the previous and the next element.

Word hyphenation is an individual writing feature covered by punctuation analysis in the forensic linguistic. People use different marks for splitting words between lines (-,-/-, =, =/=). Reconstruction of split words is crucial for automatic text processing. TEI `<hyph>` tag is used to encode word hyphenation (as in the lexicons) and the hyphen occurrence, e.g. "*competi*`<hyph>-</hyph><lb/>`*tor*". A structure of a sample suicide note is presented below (we cannot present the original scan of the letter due to the law reasons).

```
<text>
  <body>
    <pb facs="0003.1-1.png" n="1"/>
    <opener>
      <p rend="step-left"><salute>MAMUSIU TATUSIU
      KOCHANY XXXXXXXX</salute></p>
    </opener>
    <p rend="step-center">PRZEPRASZAM WAS<lb/>
      ZA TO!</p>
    <p rend="center">KOCHAM WAS BARDZO</p>
    <closer>
      <p rend="center"><signed>WASZ XXXXXX</signed></p>
    </closer>
  </body>
</text>
```

4 Case Study

Division into sentences is important for forensic linguistics, however, due to the problematic recognition of sentences limits in the notes (e.g. defective punctuation), we decided to focus on lines and paragraphs only. The way of isolating paragraphs can be characteristic to a given writer [12], e.g., indentation or by other elements: a line, a symbol sequence, initials etc. We applied TEI `<ornament>` – horizontal line with type *line* for true horizontal lines (see Fig. 1, left) and type *characters* for string of elements (e.g. asterisks) (see Fig. 1, right).

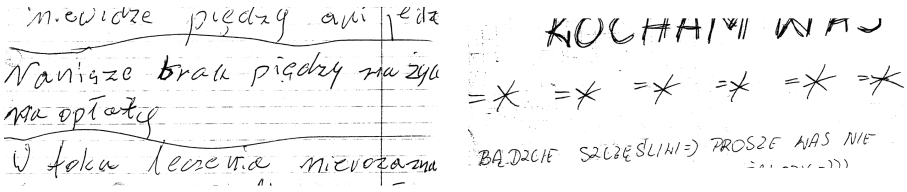

Fig. 1. Examples of horizontal lines indicating new paragraphs: (a) (left) `<ornament type="line"/>` and (b) `<ornament type="characters"/>`.

Writer's approach to hyphenation is an idiolectal feature. Authors very often avoid hyphenation that is associated with the central location of the text on the page and wide margins. Others hyphenate words in various ways, including mistakes related to syllabification ('podziekow/ać' *acknowledge*, 'pow/iedzenia' *saying*, 'króles-/stwa' *kingdom*). Wrong choice of punctuation mark or its position can be noticed, too:

- without punctuation mark ('nie/wygodni', encoded as *nie*<hyph/>
wygodni);
- with hyphenation mark at the next line beginning ('hospita/-lizacja' *hospitalization*, as *hospitaliza*
<hyph>-</hyph>*cja*);
- with double hyphenation mark ('wspom-/-ę', as *wspom*<hyph>-</hyph>
<hyph>-</hyph>*ę*);
- with equal sign ('chcia=/łem' wanted, encoded as *chcia*<hyph>=</hyph>
łem).

We extended TEI description of the hyphenation information about the usage of the punctuation mark and its location.

5 Corpus Statistics and Availability

Polish Corpus of Suicide Notes (PCSN) consists of 619 documents from years 1999–2008 obtained from prosecutor offices all over Poland. Demographical data of writers: male — 456, female — 160. The youngest authors are below 19 years old — 83 letters, the oldest were above 80 — 10 letters. Most of the suicide notes are handwritten — 604 letters, some are typed (computer, mobile phone) — 14 letters. Each note was scanned and transcribed. Fig. 2 presents the detailed statistics of corpus elements (state on the day of the 4th July 2011).

Number of	Tag	Count
documents	<body>	433
pages	<pb>	621
paragraphs	<p>	1767
line breaks	<lb/>	5427
envelopes	<envelope>	17
letter openers	<opener>	164
letter closers	<closer>	240

Number of	Tag	Count
post scriptums	<ps>	84
signatures	<signed>	197
corrections	<corr>	2996
figures	<figure>	104
hyphenations	<hyph>	90
ornaments	<ornament>	64

Fig. 2. Statistics of major elements in CPNS.

The corpus will be available to other researchers on the basis of a free research license after signing an appropriate agreement. More information about the license conditions will be published on the PCSN web page: http://pcsn.uni.wroc.pl.

6 Conclusions and Further Research

Suicide notes are short texts and a significant part of information is expressed by the note visual structure or graphical symbols. Thus, a rich annotation scheme

was proposed on the basis of TEI standard. The described aspects were divided generally into structural and related to the note content. The description of the note logical structure follows TEI adaptation for handwritten letters. As author-generated language errors can be important feature of information concerning the given individual, errors and their correction received especial attention. The proposed annotation scheme was tested on the basis of selected transcribed documents of different types. This work is continued. We plan to extend the corpus annotation with linguistic features in a semi-automated way: first applying language tools (e.g. a morpho-syntactic tagger or Named Entity recogniser) and next correcting the results manually.

Acknowledgements Research paper financed partially with funds earmarked for research project in the budget for 2010–2012.

References

1. Blackwell S.: Why Forensic linguistics Needs Corpus Linguistics. Comparative Legilinguistics 1, pp. 5–19 (2009)
2. E.S., Farberow N.L. (Eds.): Clues to Suicide, New York-Toronto-London (1957)
3. Stone P. J., Dunphy D. C., Smith M. S., and Ogilvie (Eds.): The General Inquirer: A Computer Approach to Content Analysis, Cambridge, MA: MIT Press, pp. 527–535 (1969)
4. Jones N.J., Bennell C.: The Development and Validation of Statistical Prediction Rules for Discriminating Between Genuine and Simulated Suicide Notes. In: IASR, 11 p.230 (2007)
5. Pestian J., Nasrallah H., Matykiewicz P., Bennett A., Leenaars A.: Suicide Note Classification Using Natural Language Processing: A Content Analysis, pp. 19-28, http://www.la-press.com
6. Eisenwort B., Berzlanovich A., Willlinger U., Eisenwort G., Lindorfer S., Sonneck G., Abschiedsbriefe und ihre Bedeutung innerhalb der Suizidologie., Nervenarzt, 77: p. 1359 (2006)
7. Olsson J.: Wordcrime. Solving Crime Through Forensic Linguistics, London - New York, p. 55 (2009)
8. Razak Z., Zulkiflee K., Mohd Yamani Idna Idris, Emran Mohd Tamil Mohd Noorzaily Mohamed Noor, Rosli Salleh, Mohd Yaakoob,Zulkifli Mohd Yusof, Mashkuri Yaacob: Off-line Handwriting Text Line Segmentation: A Review. In: IJCNS, vol. 8, No 7, p. 12 (2008)
9. Bosma, W., Vossen, P., Soroa, A., Rigau, G., Tesconi, M., Marchetti, A., Monachini, M. and Aliprandi, C.: KAF: a generic semantic annotation format. In: Proc. of the 5th Inter. Conf. on Generative Approaches to the Lexicon GL 2009, Pisa, Italy, (2009)
10. Vanhoutte, E., Van den Branden, R.: Describing: Transcribing, Encoding, and Editing Modern Correspondence Material: A Textbase Approach. In: Lit Linguist Computing, 24(1), 77–98 (2009)
11. Coulthard M.: On the Use of Corpora in the Analysis of Forensic Texts. In: INT J SPEECH LANG LA 1, 27–43 (1994)
12. Olsson, J.: Forensic Linguistics. An Introduction to Language, Crime and Law, London - New York, p. 52 (2007)