A WORKBENCH FOR DOCUMENT PROCESSING

Karola Witschurke

Abstract: During the MEMORIAL project time an international consortium has developed a software solution called DDW (Digital Document Workbench). It provides a set of tools to support the process of digitisation of documents from the scanning up to the retrievable presentation of the content. The attention is focused to machine typed archival documents. One of the important features is the evaluation of quality in each step of the process. The workbench consists of automatic parts as well as of parts which request human activity. The measurable improvement of 20% shows the approach is successful.

Keywords: Document Management, Digital Document Workbench, Image Processing, OCR, Machine Typed Document.

ACM Classification Keywords: I.7.5 Document and Text Processing: Document Capture

Introduction

A strategic goal of the international consortium undertaking this project was to support the creation of virtual archives based on documents which exist in libraries, archives, museums, memorials and public record offices, in order to allow computer aided information retrieval.

Machine type written documents have proved as especially hard to process.

Such documents may constitute less or more complex printed forms mixing printed and typed text, graphics, as well as hand written annotations, signatures, rubber stamps and photographs. Moreover, the colour of typed characters may vary; characters may be overstricken, shifted up or down, or due to torn out or dried ribbon only partially typed. A special challenge are documents which represent a carbon copy instead of an original one, Finally, due to physical conditions a document may contain stained or damaged parts.

1. Approach to the Project

OCR systems may be used to extract the textual information from a document image.

OCR is used to working on binary images and assigns groups of black pixels to patterns of characters. The results of state-of-the-art OCR systems are satisfying if fresh printed office documents are processed. Historical documents in general look bad for different reasons.

The problem is to get good looking binary images from bad looking coloured original ones by filtering noise (stains, wrinklings, torns) out of the image and improving the shapes of characters. Due to the unsteady quality of a page it is not satisfactory to use the same threshold of binarisation (a parameter between 1 and 256 used for binarisation of a gray scale image and causes the assignment of pixel to be “white” or “black”) for the whole page. For those reason in the MEMORIAL-project a semantic driven approach was preferred. A special editor supports the user to draw regions of interest and mark other regions with damages or illegible parts. After that a partial image improvement and background clearing using the colour information is possible. Thus, the binarisation with the best possible threshold is processed adaptively for the different regions down to single characters. The subsequent figures are illustrating the advantage of this approach.

The often automatically chosen binarsation threshold of 128 in this case provides a nearly blank image caused by the weakness of the typed characters. Figure 2 compares three manually chosen thresholds with the adaptively applied threshold of DDW. The marked region in figure 3 shows the advance of DDW vs. the best manual binarisation. In this case, the OCR is getting the best possible input.

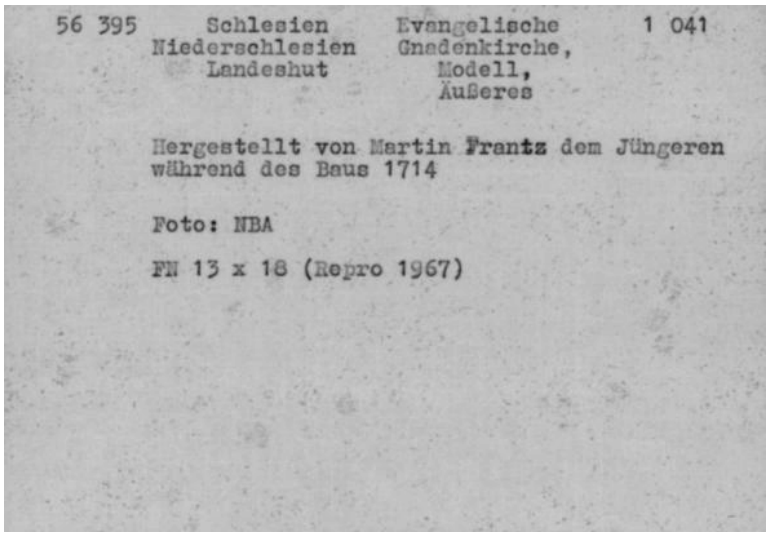

Figure 1 – inventory card from Herder-Institute Marburg (Germany)

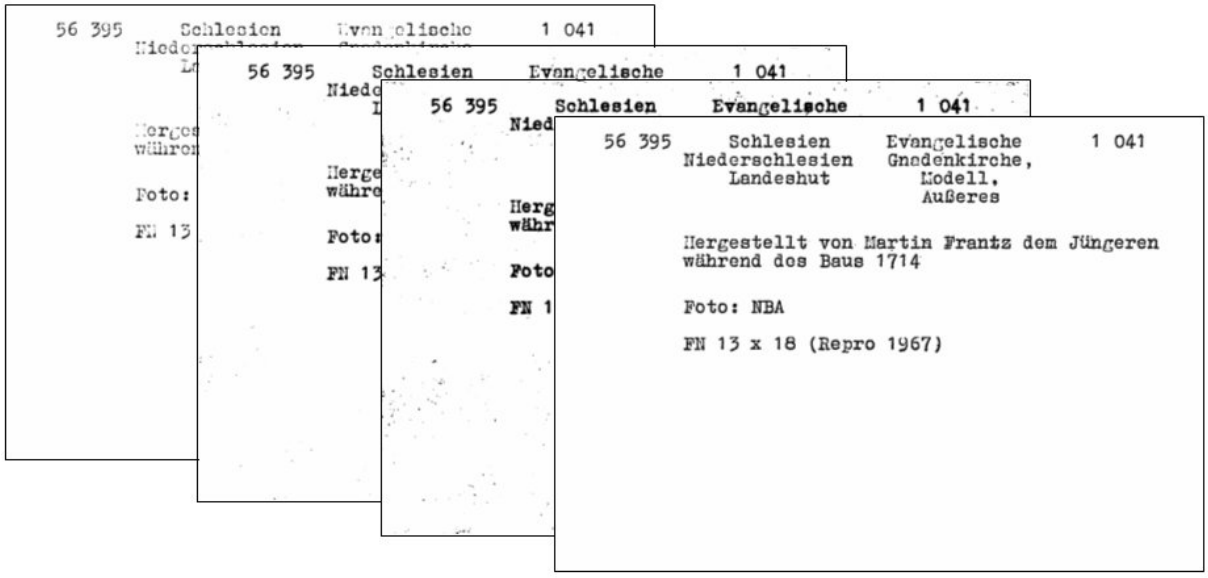

Figure 2 – comparisation of threshold (from left) 180, 200, 210 with adaptive threshold used by DDW (right)

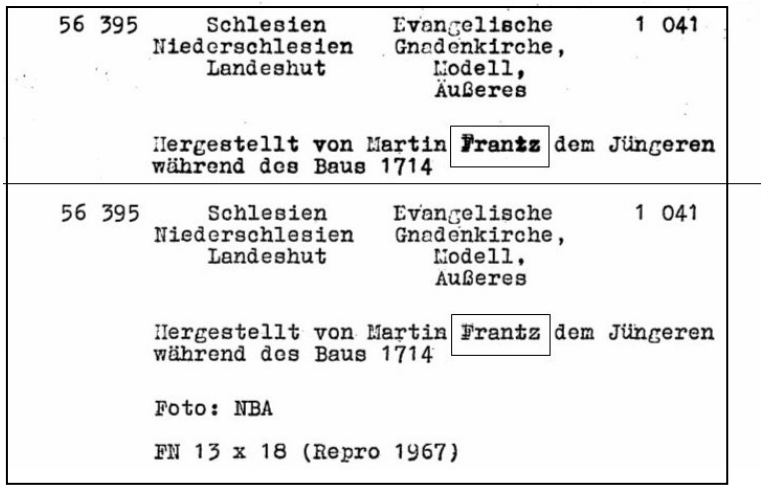

Figure 3 – best manual binarisation (above) versus adaptive thresholding

2. The DDLC Model

The transformation of a historical paper document into its electronic parallel is a complex multi-phased engineering process which may be interpreted as a document life cycle. In this effect, the project consortium has developed a *Digital Document Workbench (DDW)* toolkit supporting a *Digital Document Life-Cycle Development (DDLC)* model.

The first phase of the DDLC model is *digitisation*, which provides a raw digital image of a paper original. Depending on the constitution of the documents this process may be done manually or semiautomatically. The quality in this phase immediately influences the quality of the following phases. Furthermore the naming of the image files happens in this phase (mostly by the scanner) – each generated file must have a unique name to avoid processing duplicates or overwriting files during their processing later on. A document *Repository Management Tool (RMT)* of the DDW toolkit has been developed to help the archivist in the namespace management and to store images in a database.

The second phase of DDLC is *qualification.* An expert user should attentively classify documents by building groups of similar in structure and meaning documents. Documents within the same semantic class can be processed together throughout the rest of the cycle. The output of this phase is a XML structure called *document template*, which contains the formal description of a semantic class.

The third phase of DDLC is *segmentation*, where the identification of major components (regions) of a document page image happens. The document template is used to control the segmentation phase, where the raw document image is cleaned and improved by the *Image Processing Tool (IPT)* which transforms original (coloured) TIFF files into binarised clean images. The output of this phase is a *document content* XML file, which represents an interface for the next steps.

The fourth phase of DDLC is *extraction*, a key phase of the DDLC. Here the clean document image is processed by OCR; i.e. the textual information of an image is transformed into computer text. In this phase, the *document content* XML file is filled with the results of OCR.

The following *acceptance* phase allows the user to decide whether the recognised text is suitable to be introduced into the target database (digital archive). Corrections should be done to deliver the essential quality for the subsequent exploitation phase. This effort is supported by the content editor *Generator of Electronic Documents (GED)*. The multivalent browser *Viewer of Electronic Documents (VED)* futhermore enables addition of notifications. This might be required to improve the quality of the results of queries to the target database.

The graphical representation of the DDLC model with its interfaces is shown in figure 4.

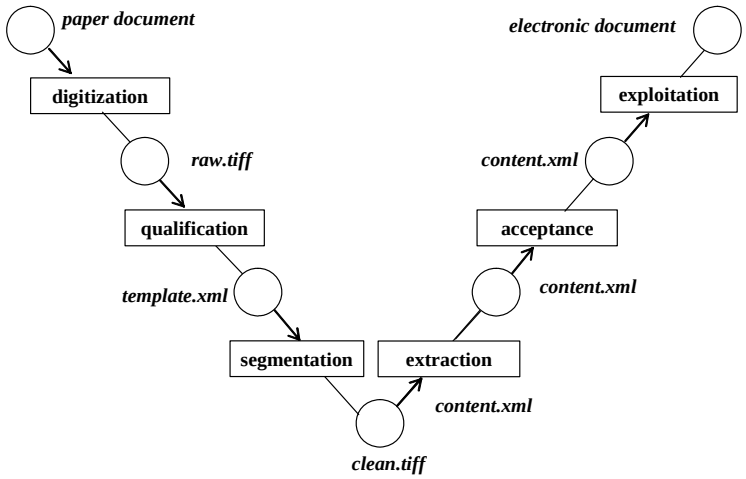

Figure 4 - DDLC model

The left part of the DDLC model represents analysis of information aided by the user, whose domain knowledge is gradually being transformed into a control structure of processes for engineering the final product, represented

by the right part. Verification of partial products of respective phases of the cycle plays a key role in assuring the quality of the final product. DDW tools enable a great deal of flexibility in extracting content of scanned paper documents into electronic documents. A sample representative subset of documents constituting a class can be processed in a *semiautomatic mode* with DDW tools to find the best settings of parameters for each DDLC phase, and next the remaining documents of the same class can be processed automatically in a batch mode, with the same settings. This approach introduced by the consortium enables fine-tuning of a document content extraction processes and quality management throughout the entire cycle.

3. Quality Management

In order to assure the best possible quality in each step of the DDLC model a quality management has been established. Therefore, human expertise is requested. The QED Tool of the DDW supports the fine-tuning of the process displaying the parameters and metrics on the one hand and setting up weights on the other hand. The interface for quality data exchange is the *qed.xml* file which is stored in the working repository (see figure 5 and table 1).

When the quality of the paper document is *Q(PD)* and quality of the electronic counterpart is *Q(ED)* then three relations are possible.

Q(PD) > *Q(ED)*, the final product quality has deteriorated during processing along DDLC phases;

Q(PD) $\cong$ *Q(ED)*, the final document quality has not significantly changed compared to the original;

Q(PD) < *Q(ED)*, the final document quality has been improved during processing.

The first case indicates incorrect parameter settings. The achieved electronic document is unacceptable. The second case indicates correct but not optimal parameter settings. The third case is the desirable one - indicating an increasing quality during processing along DDLC. It is possible only when an expert user has been able to successfully contribute to the document engineering processes.

Document quality assessment in any DDLC phase uses a specially developed Visual GQM (VGQM) method [9]. The VGQM method distinguishes between parameters and metrics. Parameters characterise processes of each phase, and their values may be used to control the DDW components. The values of metrics, specific to each phase, are measured to characterise the particular input and output data. The value of Q is calculated based on a quality tree and normalised to a five-grade scale, from very low, through low and medium, up to high, and very high quality. While all optimal settings for each phase are established by the quality expert, the processing of the remaining documents of the class can be performed automatically in a batch by an archivist. Any document that cannot pass the quality threshold set up by an expert may now be rejected. In dependence on the acceptance, a phase may rerun with tuned parameters. A thorough selection of acceptance criteria for each class implies that either document processing progresses to the next phase, or is of such a poor quality that it must be processed manually (retyped).

4. Digital Document Workbench

DDW is a set of tools which aids the user in transferring typed content of paper documents into a digital archive. Any realistic use of DDW is possible, if the following assumptions are valid:

– all paper documents are type written;
– originals may be yellowed, dirty and otherwise bad looking;
– many documents which match the same layout may be found in the paper archive;
– a target digital archive should contain machine readable text of the documents.

The "backbone" of the DDW is a MS-SQL database (of-the-shelf product) called a Working Repository (**WR**). It contains information on the entire lifecycle of each document. Component tools post their results into the working repository, this way information exchange is guaranteed between all tools. Figure 5 gives summary of the several tools belonging to the DDW and the ways of information exchange.

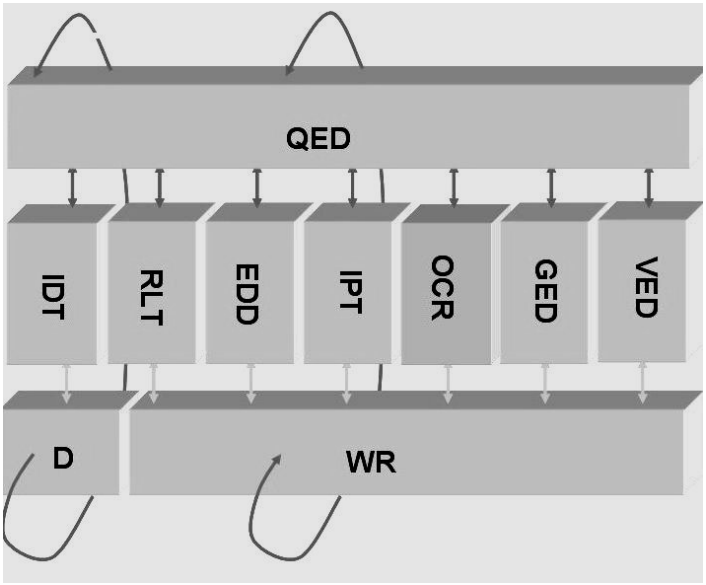

Figure 5 – Architecture of the DDW Toolset

Acronym	*Full name*	*Functionality*
IDT	InDexing Tool	Supportes name management, storage management and meta data description
RLT	Repository Loading Tool	Stores metadata and links to the image files as well as automatically generated jpeg-files and thumbnails in the working repository.
EDD	Electronic Document class template eDitor	Creates and edits a template.xml file to describe a document class, allows to add (similar) documents to the class
IPT	Image Processing Tool	Performs background cleaning and character improvement, creates a content.xml file
OCR	Optical Character Recognition tool	Separates the information relevant for OCR from the content.xml, runs the OCR (of-the-shelf product) and returns the results of OCR to the content.xml
GED	Generator of Electronic Documents	Enables editing badly recognised text or (in a extreme case) retyping a document
VED	Viewer of Electronic Documents	Allows the archivist to browse layers of the electronic document with lenses.
QED	Document Quality Evaluation Tool	Supports measurement and evaluation of the quality of each respective DDLC phase. If a quality level is not satisfactory, the whole process can be repeated with changed settings.
WR	Working Repository	Is an internal DDW database for storing project data

Table 1 – DDW components

The DDW toolkit can be used in two possible configurations:

- stand alone, without any connection to the working repository, with all relevant files stored directly in a common file system. This can be useful when just a few documents in manual mode are processed. In this case, the user has to control the file system.
- connected to the working repository (DDW database), providing a better control on intermediary document forms in between DDLC phases, in particular when operating DDW in a batch mode.

5. Final Remarks

One of the key advances of DDW is the semantic driven image processing. The success of the approach to handle colour images of documents by developed image processing tools (IPT) is shown in the following impressive graphic. It maps the confidence rate of 50 documents (register cards of Herder - Institute Marburg) depending on the threshold of binarisation. In general, OCR systems choose t =128 as threshold automatically. It can also be chosen interactively (by testing the documents and looking for the best results, here: t = 189). In both cases, the chosen threshold then holds for all documents as a whole. DDW determines the threshold value individually for each character. Figure 6 shows the enhancement of quality processing 50 different inventory cards as displayed in figure 1 with fixed thresholds vs. with an adaptively determined threshold (upper curve, resp. right column). The obtained average value of ca. 80 is not yet satisfying for the majority of archivists, but advanced methods will increase the results in further projects.

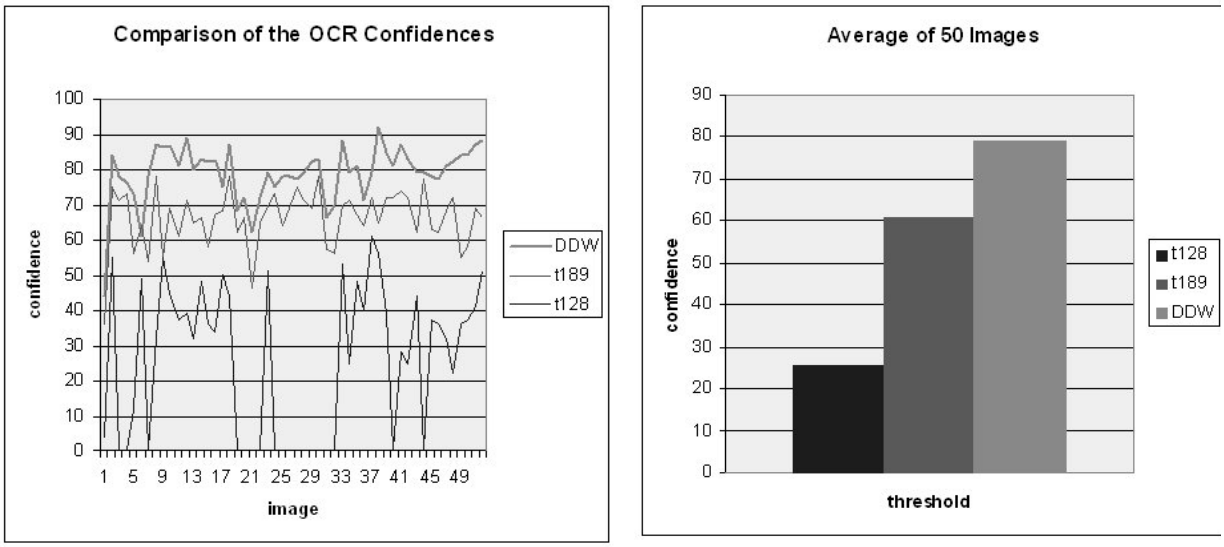

Figure 6 – Influence of binarisation parameters on the OCR results

During two workshops archivists could act as model users and test DDW with a selected set of documents. A typical situation observed during the test phase is the quality improvement combined with a significant reduction of time and effort in editing a document during the acceptance phase, compared to manual reproduction of a document from scratch.

Acknowledgements

The MEMORIAL project was funded by the European Commission within the Information Society Technologies (IST) Program from 2002-02-01 to 2004-10-31. The consortium implementing the goal statements was constituted by Gfal (Society for the Promotion of Applied Computer Science)– Germany, Moreshet Holocaust Study and Research Center– Israel, National Museum Stutthof in Sztutowo - Poland, Technical University of Gdańsk – Poland, University of Liverpool – Great Britain.

Bibliography

[1] Antonacopoulos, A., Karatzas, D., Krawczyk, H., Wiszniewski, B.; The Lifecycle of a Digital Historical Document: Structure and Content.; ACM symposium on Document Engineering, Milwaukee, USA, October 28-30, 2004

[2] A. Antonacopoulos, D. Karatzas; A Complete Approach to the conversion of typewritten Historical Documents for Digital Archives; Sixth IAPR International Workshop on Document Analysis Systems, Florence, Italy, 8-10 September 2004

[3] Szwoch M., Szwoch W.; Preprocessing and Segmentation of Bad Quality Machine Typed Paper Documents; Sixth IAPR International Workshop on Document Analysis Systems, Florence, Italy, 8-10 September 2004

[4] Apostolos Antonacopoulos; Document Image Analysis for World War II Personal Records; International Workshop on Document Image Analysis for Libraries, Palo Alto, Jan 2004

[5] Wolfgang Schade, Karola Witschurke, Cornelia Rataj; Improved character recognition of typed documents from middling and lower quality based on application depending tools Processes, results, comparison; Electronic Imaging Events in the Visual Arts - EVA 2003, Berlin 2003
[6] Alexander Geschke, Eva Fischer; Memorial Project - A complex approach to digitisation of personal records; Electronic Imaging Events in the Visual Arts - EVA 2003, Berlin 2003
[7] Henryk Krawczyk, Bogdan Wiszniewski; Definition gleichartiger Dokumententypen zur Verbesserung der Erkennbarkeit und ihre XML-Beschreibung; Electronic Imaging Events in the Visual Arts - EVA 2003, Berlin 2003
[8] Henryk Krawczyk, Bogdan Wiszniewski; Digital Document Life Cycle Development; International Symposium on Information and Communication Technologies, ISICT 2003, Dublin, Ireland
[9] Henryk Krawczyk, Bogdan Wiszniewski; Visual GQM approach to quality-driven development of electronic documents; Second International Workshop on Web Document Analysis, WDA2003, Edinburgh, UK
[10] Bogdan Wiszniewski; Projekt IST-2001-33441-MEMORIAL: Zestaw narzędziowy do tworzenia dokumentów cyfrowych z zapisów osobowych; I Krajowa Konferencja Technologii Informacyjnych 2003 TUG
[11] Alexander Geschke; MEMORIAL Project Overview; Proc. EVA Harvard, Symposium about Collaboration of Europe, Israel and USA, Harvard Library 1-2.10.2003
[12] Jacek Lebiedź, Arkadiusz Podgórski, Mariusz Szwoch; Quality Evaluation Of Computer Aided Information Retrieval From Machine Typed Paper Documents; Third conference on Computer Recognition Systems KOSYR'2003
[13] Witold Malina, Bogdan Wiszniewski; Multimedialne biblioteki cyfrowe; Sesja 50-lecia WETI-PG
[14] Dr. Alexander Geschke, Dr. Wolfgang Schade; The EU Project Memorial - Digitisation, Access, Preservation; Electronic Imaging Events in the Visual Arts - EVA 2002, Berlin 2002
[15] S. Rogerson, B. Wiszniewski; Legislation and regulation: emphasis on European approach to Data Protection, Human Rights, Freedom of Information, Intellectual Property, and Computer Abuse; PROFESSIONALISM IN SOFTWARE ENGINEERING PSE'03

Author's Information

Karola Witschurke - GFaI, Rudower Chaussee 30, 12489 Berlin, Germany; e-mail: witschurke@gfai.de

EXPERIMENTS IN DETECTION AND CORRECTION OF RUSSIAN MALAPROPISMS BY MEANS OF THE WEB

Elena Bolshakova, Igor Bolshakov, Alexey Kotlyarov

Abstract: Malapropism is a semantic error that is hardly detectable because it usually retains syntactical links between words in the sentence but replaces one content word by a similar word with quite different meaning. A method of automatic detection of malapropisms is described, based on Web statistics and a specially defined Semantic Compatibility Index (SCI). For correction of the detected errors, special dictionaries and heuristic rules are proposed, which retains only a few highly SCI-ranked correction candidates for the user's selection. Experiments on Web-assisted detection and correction of Russian malapropisms are reported, demonstrating efficacy of the described method.

Keywords: semantic error, malapropism, error correction, Web-assisted error detection, paronymy dictionaries, correction candidates, Semantic Compatibility Index.

ACM Classification Keywords: I.2.7 [Artificial Intelligence]: Natural language processing – Text analysis

Introduction

Modern computer text editors and spellers readily detect spelling errors and some syntactic errors, primarily, mistakes in word agreement. Step by step, editing facilities of computers are being extended, in particular, by