UDK 811.161.2

M. Lobur, M. Romanyshyn, A. Romaniuk
Lviv Polytechnic National University,
Computer-Aided Design Department

SENTIMENT-ANNOTATED CORPUS OF REVIEWS IN UKRAINIAN

This paper describes an attempt of creating a sentiment-annotated corpus of reviews in Ukrainian.

Key words: sentiment analysis, sentiment-annotated corpus, tools for manual text annotation.

Описано створення сентимент-анотованого корпусу відгуків українською мовою.

Ключові слова: емоційно-смисловий аналіз, сентимент-анотований корпус, засоби для анотування текстових повідомлень.

1. Problem

Sentiment analysis is the task of natural language processing, which is widely used nowadays in such areas as sociology (e.g. collecting data from social networks about people's likes and dislikes), political science (e.g. collecting data about political views of certain social groups), marketing (e.g. creating ratings of products/companies/people), medicine and psychology (e.g. detecting signs of psychological illnesses or depressive feelings in users' messages, detecting bullies with the help of messages in microblogs like Twitter), etc. [1].

There is no available sentiment analysis system for Ukrainian language, and implementing one involves having a sentiment dictionary or a sentiment-annotated corpus. Unfortunately, neither is available for Ukrainian language. This paper describes our approach to creating a sentiment-annotated corpus of Ukrainian reviews. Such a corpus can become the basis for a sentiment dictionary, a means for testing the work of a sentiment analyzer, and also a source of contexts that author can use to express his attitude toward a particular object.

2. Recent Research Analysis

The previous decade showed a rising interest in the area of sentiment analysis. This can be proven with a large number of projects, which appear every day: sentiment analysis of hotel reviews [7], bank reviews [3], restaurant reviews, comments on movies [12], products, messages about political events in blogs and social networks, etc. A big number of studies are dedicated to sentiment analysis of messages in microblogs.

With the development of interest in the ways of automatic determining of the subjectivity of a text message, a lot of supplementary tools have also been given a though to; these are different kinds of sentiment-annotated corpora, dictionaries, ontologies, lexical-semantic networks. Using sentiment-annotated corpora for sentiment analysis is not a novelty. Some detailed information on the usage of sentiment-annotated corpora can be found in the following theses on Corpus and Sentiment Analysis [2] and Sentiment in Japanese: A Corpus-Based Approach with Socio-Linguistic and Cross-Lingual Implications [5]. Current approaches to the creation of sentiment-annotated corpora can also be found in [6; 7; 8].

3. Research Aims

The aim of this research is to create a sentiment-annotated corpus of reviews in Ukrainian.

The objectives of the research are the following:

- research the task of sentiment analysis and prove the need for creation of sentiment-annotated corpus;
- analyze the available tools for manual annotation of text messages;

- develop an annotation scheme;
- analyze the results of annotating text messages.

4. Main Part

4.1. The Task of Sentiment Analysis

Sentiment analysis, or opinion mining, is a kind of text analysis, which aims to identify emotional attitudes or subjective judgments of the author concerning a particular object in the text message. The main objective of sentiment analysis is the automatic evaluation of a particular object (a person, a message in media, an event, an organization, etc.) in a text message in order to get a numerical or categorical indicator of general subjective attitude to the object. The aim of sentiment analysis is to understand the opinions and preferences of users, customers or clients.

The tasks of sentiment analysis include dynamic sentiment analysis (defining the subjectivity of text messages about the object in real time), visual sentiment analysis (graphic representation of people's attitudes to a specific object), deep sentiment analysis (a detailed analysis of sentiment values of text fragments in a review), sentiment analysis online, multilingual sentiment analysis (defining subjectivity of text messages written in different languages), etc.

This paper dwells upon deep sentiment analysis, the characteristic feature of which is that the focus of the research is not on the object of attitude, but on the text message itself. Here the aim is not simply to obtain the subjective opinion about the object, but the detailed analysis of positive, negative and neutral text fragments in the review, and, if possible, determine specific emotions that the author voices in the text message.

To implement deep sentiment analysis a lot of additional tools are used: part of speech tagging, parsing, defining lexical-semantic relations (synonymous and antonymous, in particular), the analysis of associations, machine translation systems (used for multilingual sentiment analysis), etc. A variety of dictionaries (mostly synonymous, antonymous, and sentiment), semantic networks, ontologies (like WordNet http://wordnet.princeton.edu/ or SentiWordNet http://sentiwordnet.isti.cnr.it/), and thesauri are used, too [9].

The main means of testing the work of a deep sentiment analysis system is a sentiment-annotated corpus.

4.2. Sentiment-Annotated Corpus

A sentiment-annotated corpus is a corpus of text messages, where every message is assigned a sentiment that it conveys. Creating a sentiment-annotated corpus is an integral part of the implementation process of a sentiment analysis system. There are several reasons for that.

Firstly, such semiautomatically annotated corpus provides an understanding of how people express their attitude to a particular object with the help of text (emotive language, ideograms, like smiling faces, punctuation).

Secondly, such a corpus may become the basis for a sentiment dictionary. A sentiment dictionary is a dictionary of subjective words and phrases, where each word or phrase is assigned a certain sentiment (positive, negative or neutral, if you take the broad sense; and joy, anger, pleasure, fear, etc. in a more narrow sense). Creating a sentiment dictionary from scratch is extremely time-consuming, and a sentiment-annotated corpus can become a good basis, which will save time at the initial stage of creating such a dictionary.

Thirdly, an available sentiment-annotated corpus is an excellent means of testing the developed system of sentiment analysis.

The process of creating a sentiment-annotated corpus may be divided into the following steps:

- collecting text messages for the future corpus;
- defining software for manual text annotation;
- developing an annotation scheme;
- annotating collected text messages.

4.2.1. Collecting Text Messages for the Future Corpus

In order to develop a system of deep sentiment analysis, the domain and the type of text messages that are going to be analyzed have to be defined first. Given that the subjective opinion is often expressed in the comments and reviews, it was decided to choose the review genre for the future corpus. When choosing the domain, we preferred to take restaurant reviews. This topic is relevant, as long as a lot of discussions on this topic can be observed on forums and social networks.

Restaurant reviews in Ukrainian, which became the basis of the corpus, were taken from a popular forum http://posydenky.lvivport.com/ and a website on all kinds of reviews http://v.lviv.ua/. These websites were chosen because of the big number of reviews that meet the chosen topic, and because the majority of the reviews on these websites were written in Ukrainian. This helps us to partially avoid the problem of filtering messages on a language basis and focus on creating the corpus. The structure of reviews on both websites is similar and meets the requirements of the sentiment-annotated corpus, as each review includes its author's identifier, the time, when the review was written, and the message, which contains the author's attitude.

4.2.2 Defining Software for Annotating Text Messages

There are a lot of convenient tools for manual text reviews annotation. Among the most common tools there are: Callisto, WordFreak, GATE, BRAT, DOMEO, CLaRK, Ellogon, UAM and others [10]. Let us briefly dwell upon the features of the abovementioned tools in order to define an optimal tool for our task:

Callisto (http://callisto.mitre.org) is a simple annotation tool, designed to support linguistic annotation of texts for any language that supports Unicode. Annotated texts are stored in ATLAS format, which can be easily imported into xml.

WordFreak (http://wordfreak.sourceforge.net) is a tool that supports manual and automatic annotation of linguistic data, and allows for automatic learning to correct manually-made annotations. This tool is mainly used to check the already annotated text.

GATE (http://gate.ac.uk/) is an environment for natural language processing, which also has a tool for manual and automatic text annotation. GATE provides an opportunity to create a variety of annotation schemes.

BRAT (http://brat.nlplab.org/) is an online environment for collective manual text annotation. This system was designed to handle structured data that can be processed automatically, rather than unstructured data.

DOMEO (http://annotationframework.org/) is an online environment that allows annotating texts based on an integrated ontology. This tool supports manual, semiautomatic and automatic annotation.

CLaRK (http://www.bultreebank.org/clark/index.html) is a corpus-development system, whose primary goal is to minimize manual work during the creation of linguistic resources.

Ellogon (http://www.ellogon.org/) is an open-source multilingual environment for natural language processing, which is used by individual scientists, as well as by companies engaged in the creation of natural language processing systems.

UAM (http://www.wagsoft.com/CorpusTool/) is an environment for annotating corpora, which has an imbedded feature of corpus search, as well as a graphical editor for creating annotation schemes.

Having examined the abovementioned tools for manual, automatic and semiautomatic text annotation, we found that the most suitable software for our task would be GATE, CLaRK, Ellogon or UAM. Of all these systems we chose GATE, since this system is easy to use, it provides convenient tools for editing annotated texts, gives the ability to create complex annotation schemes, allows you to store annotated texts in xml format, provides an opportunity to work with multiple text files and multiple annotation schemes at the same time, it supports Ukrainian, and also provides highlighting of annotated text with different colours, which is convenient when annotations overlap.

4.2.3 An Annotation Scheme

An annotation scheme for the sentiment-annotated corpus was designed with the help of CREOLE package (Collection of Reusable Objects for Language Engineering), which possesses an

AnnotationSchema class. This package allows you to create annotation schemas and dialogs to work with them. The configuration file creole.xml contains information about the resources that are used. In our case it is a file name with the labels that are going to be used. [10]

The developed annotation scheme for Ukrainian restaurant reviews has the following structural units:

- nickname;
- date;
- review;
- citing;
- sentence;
- clause;
- target;
- word;
- url-address.

The author of the review is marked with the label ‘nickname’. In the input data the author is indicated in the first row.

The date of the review usually follows the author.

The review itself is annotated without the citation part, if such exists. This is necessary in order to determine the subjective attitude of the author of this review, and not the author of the previously written quoted message.

The message of the previous author, which is cited in the given review, is defined with the help of the keyword «Цитата:», and labeled as ‘citing’.

Every sentence and every clause in the review are annotated separately. In the case of simple sentences, these two labels overlap. Every clause is assigned a sentiment value: positive, negative or neutral. However, the whole sentence is not assigned a sentiment value, as a complex sentence may contain both negative and positive sentiments.

If such exists, the name of the restaurant that is the object of the author’s attitude is marked too, and labelled as a ‘target’.

Every word or phrase in a subjective clause, meaning the one that has either positive or negative connotation, is labelled as a ‘word’. Each such word has a set of attributes, which have to be specified. These are:

- the lemma of the word: the value has to be written manually. This attribute is needed for the future sentiment dictionary;
- part of speech, which has the following values:

n – noun,
v – verb,
adj – adjective,
adv – adverb,
pro – pronoun,
con – conjunction,
pre – preposition,
par – particle,
exc – exclamation,
num – number,
und – undefined (e.g., smiling faces).

The part of speech is needed in order to differentiate the sentiments of the homoforms in the sentiment dictionary;

- sentiment, which has the following values:

positive – words and phrases that convey positive attitude;
negative – words and phrases that convey negative attitude;

neutral – words and phrases that do not convey neither positive, nor negative attitude;

intensifier – words-amplifiers that do not convey an independent attitude, but enhance the sentiment of the next word or clause. These are such words as «дуже», «надзвичайно», «безмежно», «вкрай», «досить»;

invertor – words-invertors, which do not have an independent sentiment, but change the sentiment of the next word or clause to the opposite. These are such words as «не», «нема», «немає», «неможливо», «нереально», «ніяк»;

- emotion, which has the following values: joy, sadness, anger, fear, disgust, surprise, and none (if a word or a phrase do not convey any specific emotion).

The attributes 'sentiment' and 'emotion' will have the values 'neutral' and 'none' for conjunctions, pronouns, prepositions, particles, exclamations, numbers and undefined words. Only nouns, verbs, adjectives and adverbs may have other values for those attributes.

To determine the set of basic emotions for our annotation scheme, we analyzed sets of basic emotions developed by different psychologists (see Table 1). Six of the abovementioned emotions are considered to be basic human emotions, according to the theory of a renowned psychologist Paul Ekman. Basic emotions by P. Ekman are culturally independent emotions that every person obtains during the first six months of his life. It is also a set of emotions that are easily expressed with the help of mimicry and verbal means [4]. Of course, there are other views on this issue, but basic Ekman emotions are considered to be standard.

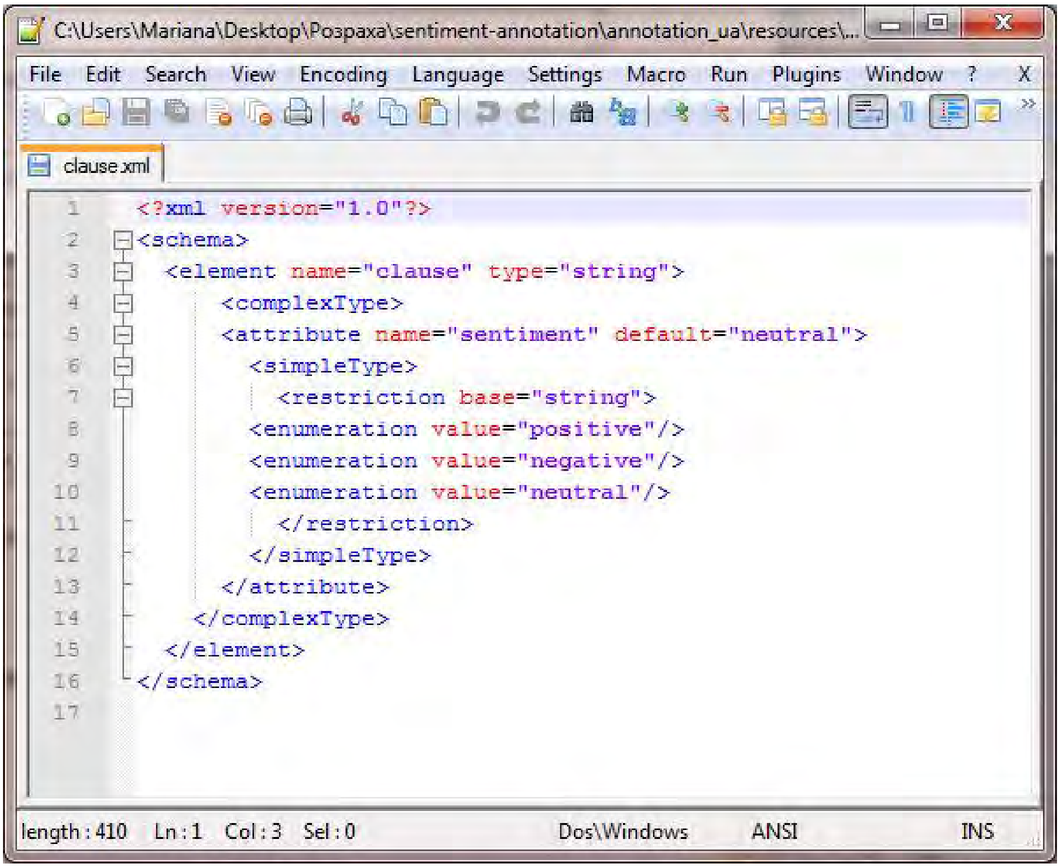
The url-addresses are labeled, too.

Table 1

Basic emotions

Psycologist	Basic emotions
R. Plutchik	Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise
M. Arnold	Anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love, sadness
P. Ekman, V. Friesen and F. Ellsworth	Anger, disgust, fear, joy, sadness, surprise
N. Frijda	Desire, happiness, interest, surprise, wonder, sorrow
J. Gray	Rage and terror, anxiety, joy
K. Izard	Anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise
V. James	Fear, grief, love, rage
V. McDougall	Anger, disgust, elation, fear, subjection, tender-emotion, wonder
O. Mowrer	Pain, pleasure
K. Oatley and F. Johnson-Laird	Anger, disgust, anxiety, happiness, sadness
J. Panksepp	Expectancy, fear, rage, panic
S. Tomkins	Anger, interest, contempt, disgust, distress, fear, joy, shame, surprise
J. Watson	Fear, love, rage
B. Weiner and G. Graham	Happiness, sadness
J. W. Parrott	Love, joy, surprise, anger, sadness, fear

The information about each label of the designed annotation scheme is written in a separate xml-file of a certain structure. Figure 1 shows the file structure for the label 'clause'.

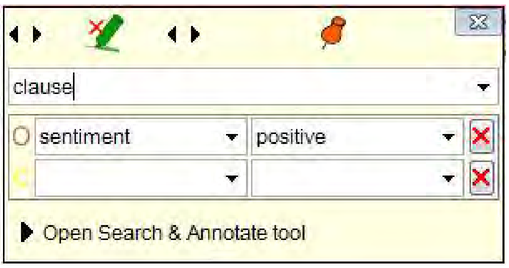

C:\Users\Mariana\Desktop\Pospaxa\sentiment-annotation\annotation_ua\resources\...
File Edit Search View Encoding Language Settings Macro Run Plugins Window ? X
clause.xml

```
<?xml version="1.0"?>
<schema>
  <element name="clause" type="string">
      <complexType>
      <attribute name="sentiment" default="neutral">
        <simpleType>
          <restriction base="string">
        <enumeration value="positive"/>
        <enumeration value="negative"/>
        <enumeration value="neutral"/>
          </restriction>
        </simpleType>
      </attribute>
    </complexType>
  </element>
</schema>
```

length : 410 Ln : 1 Col : 3 Sel : 0 Dos\Windows ANSI INS

Fig. 1. Structure of the clause.xml file

Figure 2 shows a dialog box in the GATE 7.0 environment with the label’s name, attribute and attribute value. The value of the attribute does not have to be entered manually, as they can be selected from the drop down list.


clause

sentiment positive

Fig. 2. A dialogue box for a label ‘clause’
in the Gate 7.0 environment

An appropriate part of code about the label has to be inserted into the configuration file creole.xml:

```
<AUTOINSTANCE>
  <PARAM NAME ="xmlFileUrl" VALUE ="resources/schema/clause.xml" />
</AUTOINSTANCE>
```

4.2.4 Annotating the Collected Text Messages

In order to create a sentiment-annotated corpus of reviews in Ukrainian, master students of the applied linguistics department were attracted. There was a course paper developed within the course of

Computational Linguistics. The aim of the paper was to introduce the practice of creating corpora and manual annotation of text messages to students.

Every student received a guide on the GATE environment, a detailed description of the annotation scheme with explanations and examples, as well as a set of restaurant reviews in Ukrainian in the txt format. In order to be objective, every review was given to two students for annotating. In that way we got it possible to choose a review that was better annotated.

After annotating the reviews students saved them in an xml format.

The result of the course paper was a sentiment-annotated corpus of reviews in Ukrainian, saved in an easy to use format. This corpus, however, required some additional verification because of a significant number of errors. Only after such verification the corpus will be ready to use for future research.

4.3. An Example of an Annotated Review

Let us provide a specific example:

Artemida
07.05.2011, 16:42
Всім дуже сподобалось у "Герольді", але погоджуюсь з Танею, що ціни там не з дешевих.

A review is composed of an author identifier, date and time, when the review was written, and the message itself.

After being annotated, the review was saved in xml format. The structure of xml-file contains information about encoding, information about the document itself, the review with the borders of annotations and, finally, the descriptions of the labels.

The review with the borders of annotations has the following structure:

```
<TextWithNodes><Node id="0" />Artemida<Node id="8" />&#xd;
&#xd;
<Node id="12" />07.05.2011, 16:42&#xd;<Node id="30" />
&#xd;
<Node id="124" />Всім<Node id="128" /> <Node id="129" />дуже<Node id="133" />
<Node id="134" />сподобалось<Node id="145" /> <Node id="146" />у<Node id="147" /> "<Node
id="149" />Герольді<Node id="157" />", <Node id="160" />але погоджуюсь з Танею<Node
id="182" />, <Node id="184" />що<Node id="186" /> <Node id="187" />ціни<Node id="191" />
<Node id="192" />там<Node id="195" /> <Node id="196" />не<Node id="198" /> <Node
id="199" />з<Node id="200" /> <Node id="201" />дешевих<Node id="208" />.&#xd;<Node
id="210" />
</TextWithNodes>
```

An example of the ‘nickname’ label in xml format:

```
<Annotation Id="4" Type="nickname" StartNode="0" EndNode="8">
</Annotation>
```

An example of the ‘clause’ label in xml format:

```
<Annotation Id="13" Type="clause" StartNode="184" EndNode="208">
<Feature>
  <Name className="java.lang.String">sentiment</Name>
  <Value className="java.lang.String">negative</Value>
</Feature>
```

```
</Annotation>
```

An example of the ‘word’ label in xml format:

```
<Annotation Id="25" Type="word" StartNode="134" EndNode="145">
<Feature>
  <Name className="java.lang.String">part_of_speech</Name>
  <Value className="java.lang.String">v</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">lemma</Name>
  <Value className="java.lang.String">подобатися</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">sentiment</Name>
  <Value className="java.lang.String">positive</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">emotion</Name>
  <Value className="java.lang.String">joy</Value>
</Feature>
</Annotation>
```

From the above shown examples we can see that each label contains the start and end indices of the text fragment, annotated with this label, as well as the names and values of its attributes, if such exist. Such an xml-tree can be easily parsed and used for future research.

Conclusion

The process of creating a sentiment-annotated corpus can be divided into the following stages: collecting text messages for the future corpus, defining software tools for annotating text messages, development of an annotation scheme and annotating the collected text messages.

This article describes our attempt of creating a sentiment-annotated corpus of Ukrainian restaurant reviews with the help of GATE environment. The developed annotation scheme and the structure of an annotated review have been described.

1. Давыдов А. А. Системная социология: Opinion Mining / А. А. Давыдов. – М.: ИС РАН, 2009. – Режим доступу: http://www.isras.ru/index.php?page_id=1024 2. Cheng T. Corpus and Sentiment Analysis / Tai Wai David Cheng. – Guildford, 2007. – 144 ст. 3. Deep sentiment analysis with attensity analyze optimises Lloyds’ customer service. – Режим доступу: http://www.attensity.com/wp-content/uploads/2010/09/LloydsSuccessStory.pdf 4.Ekman P. Basic Emotions / Paul Ekman // Handbook of Cognition and Emotion. – John Willey & Sons Ltd, 1999. – P. 45–60. 5.Grissom A. Sentiment in Japanese: A Corpus-Based Approach with Socio-Linguistic and Cross-Lingual Implications / Alvin Castillo Grissom II. – Hendrix College, 2006. – 132 p. 6.Kabadjov M. Sentiment Intensity: Is it a Good Summary Indicator? / M. Kabadjov, A. Balahur, E. Boldrini // Human Language Technology. Challenges for Computer Science and Linguistics: 4th Language and Technology Conference, LTC 2009. – Roznan, Poland, November 6–8, 2009. – P. 205–207. 7.Kasper W. Sentiment Analysis for Hotel Reviews / Walter Kasper, Mihaela Vela. – Proceedings of the Computational Linguistics-Applications Conference. – Jachranka, Poland: Polskie Towarzystwo Informatyczne, Katowice, 10/2011. – P. 45–52. 8.Kessler J. The ICWSM 2010 JDPA Sentiment Corpus for the Automotive Domain / Jason S. Kessler, Miriam Eckert, Lyndsay Clark, Nicolas Nicolov. – 4th International AAAI Conference on Weblogs and Social Media Data Challenge Workshop (ICWSM-DCW 2010). – Washington, D.C., 2010. – 8 p. – Режим доступу: http://www.icwsm.org/2010/papers/icwsm10dcw_8.pdf 9.Santos A. Determining the Polarity of Words through a Common

Online Dictionary / Antonio Paulo Santos, Carlos Ramos, Nuno C. Marques // EPIA'11 Proceedings of the 15th Portugese conference on Progress in artificial intelligence. – Berlin, Heidelberg, 2011. – P. 649-663. – Режим доступу: http://ssdi.di.fct.unl.pt/~nmm/MyPapers/SRM2011_PublishedBySpringer_EPIA.pdf 10)Shapiro S. Natural Language Tools for Information Extraction for Soft Target Exploitation and Fusion / Stuart C. Shapiro, Shane Axtell. – NY, 2007. – P. 36–37. – Режим доступу: http://www.cse.buffalo.edu/~shapiro/Papers/shaaxt07.pdf 11. Using GATE Developer. – Режим доступу: http://gate.ac.uk/sale/tao/splitch3.html#chap:developer 12. Yessenov. Sentiment Analysis of Movie Review Comments / Yessenov, Kuat, Sasa Misailovic. – Massachusetts Institute of Technology, Spring 2009. – Режим доступу: http://people.csail.mit.edu/kuat/courses/6.863/report.pdf

UDC 004.8

V. Stupnytskyy
Lviv Polytechnic National University

SUBSYSTEM OF RHEOLOGICAL FORMING MODELLING IN INTEGRATED CAD/CAPP/CAM SYSTEM IN MACHINE BUILDING

In the article the brought analysis of automated technological planning process for the machine-building production (ICAM) trends. The offered perfection of ICAM structure of the system as introduction of the Computer Aided Forming sub-system (CAF-system). The brought arguments over in relation to introduction of conception of the parallel engineering, introduction of CALS – technologies and functionally-oriented technologies.

Key words: automated technological planning process, CALS, CAF, rheological modelling.

Описано аналіз процесу автоматизованого технологічного планування машинобудівного виробництва (ICAM). Запропоновано вдосконалення структури системи ICAM із впровадженням підсистеми автоматизованого формування (CAF-системи). Наведено аргументи щодо впровадження концепції паралельної розробки, впровадження CALS-технологій та функціонально-орієнтованих технологій.

Ключові слова: процес автоматизованого технологічного планування, CALS, CAF, реологічне моделювання.

Introduction

The generalized analysis of trends for the modern machine-building computer-assisted operation sequence planning systems (CAD/CAPP/CAM/PDM) gives a possibility to mark such features.

For all most effective machine-building CAD of middle and high level (Pro/Engineer, Unigraphics, CATIA, SolidWorks; Nastran, Solid Edge) characteristic system integration of software product (optimally is creation of hybrid CAD/CAE/CAPP/CAM software) is with the aim of more effective exchange by design-engineering information in only compatible formats and prototypes of data repository (MIL – STD – 2549 Configuration Management Data Interface). In addition, there is a tendency to unitization of technological preparation, that shows up in the use normatively-legal base of CALS-technology (ISO 11179, MIL – STD – 1840, MIL – STD – 1808A, MIL – STD – 974) and others like that.

Introduction of PLM (CALS) – technologies are required by planning of the functionally-oriented technologies for machine-building production, i.e. taking into account already on the stage of technological preproduction of not only parameters of exactness of sizes and roughness of the processed surfaces but also complex of new qualimetrycal indexes that will have substantial influence on providing of operating, repair, heat-recovery and other functional properties of the fabricated products.