Computer Science • 17 (1) 2016 http://dx.doi.org/10.7494/csci.2016.17.1.23

Agnieszka Pluwak
Wojciech Korczynski
Marek Kisiel-Dorohinicki

ADAPTING A CONSTITUENCY PARSER TO USER-GENERATED CONTENT IN POLISH OPINION MINING

Abstract *The paper focuses on the adjustment of NLP tools for Polish; e.g., morphological analyzers and parsers, to user-generated content (UGC). The authors discuss two rule-based techniques applied to improve their efficiency: preprocessing (text normalization) and parser adaptation (modified segmentation and parsing rules). A new solution to handle OOVs based on inflectional translation is also offered.*

Keywords user generated content, text normalization, parsing, sentiment analysis

Citation Computer Science 17 (1) 2016: 23–44

1. Introduction

User-generated content (henceforth, UGC) as data (text, pictures, videos) authored by Internet users has recently been facing a rapid volume growth (e.g., every minute „(...) Facebook users share nearly 2.5 million pieces of content, and Twitter users tweet nearly 300,000 times (...)" [10]). The textual type of UGC, such as tweets, Facebook posts, fora contributions, as well as *opinion reviews* (evaluation of products or services) is notorious for decreasing the performance of natural language processing systems. This can be ascribed to the complex characteristics of colloquial language used by opinion givers online (written-down spoken discourse, see *secondary orality* [15, 25]). The reason that processing (of UGC content e.g., parsing) brings poorer results is also to be attributed, among others, to the fact that statistical, corpus-based tools are trained on available national corpora with a standard makeup of mostly stylistically correct texts originating in the journalist-written media (e.g., [13, 19]), whereas UGC often does not abide by these rules. As we have identified in the example of Polish opinion reviews, the major recurrent problematic phenomena are, for example:

- **deficiencies in lexica**, such as colloquialisms (e.g., *an app* standing for *an application*), neologisms (e.g., a Polish diminutive derivative of *an [English] application* is *appka*) and abundant named entities (e.g., product names, such as *Toshiba*) often with digits in brand and model names, emphasized by emoticons, either split into separate characters by parsers or not recognized at all;
- **lack of diacritics and some orthographic or spelling mistakes**, making a known lexeme unrecognizable for the dictionary, morphological analyzer or tagger;
- **relaxed punctuation** – deficits in punctuation marks impacting sentence and phrase segmentation and general sentence structure (syntax).

The challenge of UGC text processing, especially with regard to opinion reviews on social media platforms, is therefore a widely known issue, largely debated at the 2013 ACL Conference [24, 31, 32] or 2013 Sentiment Analysis Symposium (e.g., [30]). However, only a few attempts have been made so far to process Polish UGC, partly because – in comparison with English, for example, it still is a quite resource-poor language in terms of NLP. In order to cover this demand, we would like to offer a solution based on the use of available tools. We would like to present the first attempt (to the best of our knowledge) to adapt a deep parser of Polish to deal with three common UGC problems present in opinion reviews: character correction, phrase plus sentence segmentation, and new lexeme recognition (OOVs – words out of the standard vocabulary). We also tried to improve its general performance by adding a few new complex syntactic constructions that have not been processed by the parser so far.

The hereby-presented method poses an attempt of text normalization via parser tuning and pre-processing applied to Polish opinion mining in the hotel, restaurant, laptop, and mobile application domain with an adaptation of a rule-based approach.

The makeup of the solutions results from a thorough linguistic analysis of the characteristics of opinion reviews: 1. an analysis of typical phenomena of opinion reviews influencing the parser's performance on a specimen of 260 sentences; 2. Study of share of a parser's error types conducted on a specimen of 250 opinion texts (about 1000 sentences).

Within this approach, we also have identified phenomena characteristic of a colloquial register with patterns recurrent in a variety of languages, as well as typical features of the Polish UGC language. Therefore, the solutions offered span from character to token-level, requiring fixing of the lack of characters (on the one hand) up to the replacement of unknown lexemes such as named entities (on the other). The system is layered with a Segmentation module (relaxed segmentation rules, dealing with punctuation anomalies), a Correction module (providing for correction of orthographic mistakes), and the Inflection provider, replacing unknown vocabulary (OOVs) with the known lexemes for parsing and dealing with lexical deficiencies in the default morphological analyzer.

The novelty of our approach is based on two features: 1. a rule-based modification and adaptation of the existing tools, such as the Gobio parser[1], to UGC processing instead of implementing an entire pre-processing stage or machine learning techniques; 2. use of the highly inflectional makeup of the Polish language in translation of unknown tokens. Inflection, a characteristic feature of both Polish and many other Slavic languages, has already been commented upon in the NLP literature [26] as a mixed blessing for an NLP system structure: on the one hand, it seems more laborious to work on due to a greater number of morphological cases (especially for lexicographers preparing inflectional or morphological dictionaries), and on the other – it is far less ambiguous to process than a context-dependent language like English. Therefore, the inflectional character of the Polish language enables the construction of a translation module based on the case – morphemes (suffixes) to replace the unknown entities with the synonyms of the same inflectional patterns, recognized within the dictionary.

This paper consists of the following parts: state-of-the-art of text normalization in Polish and other languages (Section 2), linguistic characteristics of Polish opinion reviews (Section 3), tests conducted on available NLP-tools for Polish (Section 4), experiments in adjustment of selected modules to UGC-processing (Section 5), test results (Section 6), and conclusions (Section 7).

2. Related work

There are several available techniques of text normalization used in NLP ranging; from partly manual, partly rule-based text correction through preparation of the training corpus [36] up to machine translation and unsupervised machine learning with the use of corpora or dictionaries [32]. Rodriguez-Penagos in [32], basing on [15], mentions *spell-checking* via Hidden Markov Models [2, 5, 6, 38], *machine translation*

[1] `http://psi-toolkit.wmi.amu.edu.pl/help/documentation.html`

with substitution dictionaries [1, 28], and an *unsupervised machine learning* (imitation of the human text correction based on letter transformation, visual priming, and string/phonetic similarity [18]) as current major approaches to this issue. Typically, methods are applied either on the token (e.g., token translation) or character levels [32].

Like this work, most of these methods share the recognition of linguistic phenomena characteristic of the UGC register, influencing the parser's performance. Although the machine-learning techniques seem to prevail in the aforementioned studies, we have been inspired by the rule-based methods of parser and POS-tagger tuning, much as fixing of Penn Treebank errors with deterministic rules as described by Manning in [20], especially for the purposes of unknown vocabulary recognition. In his study, Manning analyzes the 100 most recurrent parser errors and offers rule-based solutions to specific problems that hinder a parser's good performance, such as lexicon gaps, unknown vocabulary, difficult linguistics, or having no standard to learn from. Although, as Chiticariu et al. state in [4], the rule-based approach is often looked upon as a little bit outdated or less estimated in comparison with machine learning, it turns out to be quite efficient in the task in question. On the Polish ground, at least one such attempt is known; namely, the rule-based TAKIPI tagger[2] tuning by Piasecki [26].

The subject of UGC-normalization in the Polish context has not been given much attention so far. Due to the fact that most business approaches to sentiment analysis of social media in Polish use the bag-of-words approach (e.g., Brand24, SentiOne, Brandometr, Guarda BRD, Instytut Monitorowania Mediów, MediaMon, NewsPoint, Sentymetr) or even do not use automatic methods at all (Press-Service, see: [29]), only some of them mention selected choices of adaptation in their tool descriptions, such as slang or emoticon recognition (e.g., Brand24). Within the Polish NLP literature, two examples of text pre-processing for shallow parsing in sentiment analysis have been identified; namely, Sentipejd [3] and TrendMiner[3] [23].

The most extensive description of UGC-adaptation mechanisms for Polish can be found within the Multilanguage TrendMiner project in the case study of Polish political tweets [23]. Within this study, text normalization of Twitter-specific phenomena (such as URL, shortened utterances, limited by the number of digits in a tweet, hashtags) is performed with the use of regular expressions. The preprocessing tool performs the function of removal of (e.g.) URLs, translation of hashtag-coded phrases, @mentions or re-tweets, and recognition of emoticons. It also includes a rule-based spelling corrector called LanguageTool and performs corrections of misspellings. Incorrect-word recognition is carried out via the following: 1. comparison of tokens with available dictionary entries; 2. selection of all possible token interpretations; and 3. N-gram-based corpus search of the balanced specimen of the Polish National Corpus (Narodowy Korpus Języka Polskiego[4]), aimed at detection of the token with the

[2] http://nlp.pwr.wroc.pl/takipi
[3] http://www.trendminer-project.eu
[4] http://www.nkjp.pl

most frequent occurrence of all tokens identified in step 2. The F1 measures of the system ranged from 0.6 to 0.75.

Buczyński and Wawer [3] in the second approach do not provide specific details about the methods used. They apply some kind of a correction mechanism returning the proper lexeme spelling called "a dedicated procedure (...) to guess the missing diacritics." Finally, on the website of Applica[5], one can find a description of an architecture of a system performing the function of sentiment analysis; however, no paper or thorough explanation as to further characteristics was found. Some kind of text correction is conducted; however, there is no white paper or publication on the website explaining the technicalities in a detailed fashion.

The major differences between the TrendMiner approach to Polish UGC text normalization and the one presented here are the proportions of linguistic phenomena problematic for parsing and interference with the parser's segmentation versus pure pre-processing. The character of UGC is different in tweets and opinion texts, mostly due to the different profiles of their authors. Tweets by politicians are of a stylistically rather-correct character, enabling the use of an available sentiment dictionary (e.g., Słownik Wydźwięku), whereas opinion reviews are more complex in terms of language phenomena, less correct, and have a different share of problems (in tweets, the major issue are abbreviations – 26%, lack of Polish diacritics – about 10%, and emoticons – 12%, while loanwords and misspellings correspond to 3% of mistakes only; in opinion reviews, the major issue is the recognition of product names, brands, and other proper names, and loanwords amounting to about 50% of all mistakes). What is more, if you are aiming at an analysis spanning above the phrase level towards the sentence and paragraph levels, you need to take into consideration more factors such as punctuation, etc. – described below in Section 3. For this reason, the stress in our work is put on token translation of OOVs, diacritics correction, and improvement of the parser's segmentation rules.

3. Short linguistic characteristics of opinion reviews

As a research sample, we chose a set of 55 stylistically different texts of opinion reviews (260 sentences, 1452 words) written by the users on the following websites: Foursquare[6] (15 short opinions about the quality of restaurants in Krakow); Trip Advisor[7] (10 long opinions on the services of the Intercontinental hotel in Warsaw); Opineo[8] (10 medium-length opinions on an HP laptop); and Appstore[9] (20 short reviews concerning a mobile application used to learn how to play the piano). The texts ranged in length between one phrase-texts (e.g., two to four words, *Najlepsza czekolada w mie-*

[5] `http://applica.pl/rozwiazanie-applica/architektura-rozwiazania`
[6] `http://www.foursquare.com`
[7] `http://www.tripadvisor.pl`
[8] `http://www.opineo.pl`
[9] `http://www.store.apple.com.pl`

ście/The best chocolate in town) to elaborated 10-sentence-long opinions, reflecting the variety of writing styles and discourse created within this Internet genre.

The major feature of opinion reviews (typically omitted in NLP publications) that we paid attention to was its *secondary orality*, meaning that – as an Internet genre – it is characterized by "the use of electronics for oral communication" [25]. Contemporary changes in the means of communication have resulted in a hybrid character of the Internet discourse, which now owns the features of both written and spoken language, where the latter prevails.

A linguistic analysis helped us define the following characteristics of opinion reviews, which bear features of the written-down speech:

- On the lexical level:
  - colloquialisms and youth register: *lapek/lappy*, *nówka/brand new*, *supcio/goodish*, *git/OK*, *sprzęcik/gear*, *spoko gierka/cool app*;
  - technology-related loanwords from the English language, typically abbreviations, such as *t-pad*, *wifi* or neologisms derivative from loanwords; e.g., *appka* (*an app*);
  - new or popular abbreviations not yet included in morphological analyzers; e.g., *wg – według* (*according to*) or new abbreviations, such as *k. graficzna* for *karta graficzna* (*display card*);
  - chat, instant messaging or SMS-type language: abbreviations with digits (e.g., *2os – dwuosobowy*, pl. *two-person* (adj.)), dates, phonetic acronyms (*3ma – trzyma/ holds*), often resulting in the lack of analysis due to the automatic separation of digits and letters; emoticons and other substitutes of non-verbal communication, such as *hehe* [9, 37], frequently analyzed by the parsers as separate punctuation marks;
  - domain-related vocabulary, such as proper names of products, models, services etc.; e.g., *Ipad*, *HP nx7400*, *Roiboos*, SPA or *features* or aspects of evaluated products [12], e.g. *RAM-memory*, *touch pad*, *keyboard*, *screen*, *USB-port*, *S-Video*);
  - vulgarisms (some as loanwords); e.g., *szit/shit*.

  Parsing of a sentence including the above-mentioned vocabulary typically results in at least one error: the lack of a token analysis (in the case of the Gobio parser) or the lack of sentence analysis (in the case of other parsers). This could be attributed to deficiencies in the lexical resources of morphological analyzers.
- On the level of punctuation:
  - lack of (or poor) punctuation resulting in the lack of a coherent sentence structure, asyndetons; e.g., *wspaniała aplikacja nic dodać nic ująć – great application perfect just perfect*;
  - polisyndetons – overabundance of the same type of linking words; e.g., *and the battery is bad, and the disc crashes, and the mouse goes bad easily*;

  - lack of Polish characters (e.g., *ż, ą, ę*; e.g., *powazny blad* instead of *poważny błąd*) or even spelling mistakes (e.g., *powżany bąłd*), making the recognition of a lexeme difficult;
  - lack of space between words resulting in a change of meaning or secondary homonymy (e.g., *nie ma – there are no* – verb → *niema – deaf* – adjective);
  - exclamation or emphasis through multiple punctuation marks [9, 37]; e.g., *supcio/goodie-goodie!!!!!!*

  Since the punctuation marks often pose anchors for segmentation rules in NLP, wrong punctuation results in an immediate lack of segmentation and incorrect outcome of the parsing process.
- On the level of syntax:
  - sentences with incorrect (anacoluthic) syntactic structure [37]; e.g., *Dla mnie rewelacja, odkąd poznałam to miejsce./To me (it's been) great, since I've known this place.*
  - eliptic syntax, omission of predicates or unfinished sentences: *Boski sernik, najlepsza szarlotka w mieście./Delicious cheesecake, best apple pie in town.*
  - self-corrections; e.g., *to jest, to znaczy.../this is, this means...* Because the predicate quite often becomes the anchor point of a sentence, its omission in elliptic sentences frequently results in the parser not recognizing the sentence as a whole.
- On the level of orthography:
  - spelling mistakes; e.g., *ochyda* instead of correct *ohyda* (*disgusting*), resulting in the lack of recognition of a given word, and thus lack of analysis;
  - omission of capital letters in proper names or at the beginning of the sentence; e.g., *reksio* (a famous dog character from a Polish cartoon, an English equivalent could be Pluto, for example); *najlepsza czekolada!/the best chocolate!* [9];
- On the level of phonetics:
  - emphasis or exclamation by the repetition of characters: for example: *sssssssssssssssssssuuuuuuuuuuuuuuuppppppppppccccccccciiiiiiiooo/gggggggggrrrrrrreeeeeeeaaaaatt.*

Finally, opinion reviews are also characterized by stylistic and pragmatic features. In terms of style, we encounter apostrophes or direct references, comparisons, and irony, for example. As to pragmatics, the key feature is teleology (as opposed to dialogue; see: [37]), meaning that the author of a review refers to many different addressees simultaneously (e.g., prospective customers of a product and its producer). Since speech acts and pragmatics of opinion reviews are not the focus of this article, suffice it to say that opinion reviews tend to have different and sometimes unspecified addressees (ranging from other customer, through service-providers, and producers up to all of them included or an undefined general addressee), which results in the use of different modes (imperative, indicative, and interrogatory).

The statistics corresponding to the five major types of errors were as follows:

- **Lexical** (**50%** of all mistakes; 137 tokens out of 1452 words) understood as **unrecognized lexemes** (proper names, digits, instant messaging language, loanwords, etc.), always resulting in the lack of a proper parse tree. This factor reduced the number of correct parse trees by half, since no unrecognized lexeme is provided a tag.
- **Punctuation** (about **25%** of all mistakes; present in 19 out of 55 opinions). This group of errors has to be divided into token and sentence levels. On the sentence level, 11 out of 55 opinions included such problems as anacoluthic structure, phrase repetition, or lack of space, whereas 46 tokens were not recognized due to the lack of Polish digits (8 opinion texts). This means that parsing of over one third (**36%**) of all opinion texts was influenced by improper punctuation, resulting in improper sentence and phrase segmentation.
- **Syntactic**, meaning the lack of proper sentence structure, combining compound sentences into one, "streams of consciousness writing" (referring to at least one fourth of all sentences and causing **13%** of all errors).
- **Orthographic mistakes**, typically misspellings of lexemes (only about **5%** of all mistakes).
- **Morpho-syntactic**, meaning mistakes made by the parser, such as provision of a wrong tag (about **7-8%** of all mistakes, 21 tokens only out of 1452 words).

The total number of errors was 271. Therefore, on average, there was at least one error of any of the above-mentioned types per sentence. The following conclusions were drawn from these results:

- if the major problem was on the token level and referred to unrecognized lexemes, their recognition should greatly improve the parser's general performance (we checked via manual replacement of token from unrecognized to recognized and received correct parse trees);
- if syntactic and punctuation mistakes were related to improper sentence segmentation (lack of dots, streams of thought), by improving (e.g., relaxing rules of the segmenter general sentence segmentation, and the parser's performance should improve;
- if orthographic mistakes or misspellings make the tokens unrecognizable to the parser, a simple corrector should fix both the lack of Polish diacritics and the misspellings;
- if the parser's proper errors, apart from UGC-related mistakes (e.g., improper POS tags due to separation of compound adverbs; e.g., *for sure*) make only 7-8% of all errors, and for sentences devoid of misspellings, unrecognized lexemes, etc. (the same sentences translated into correct forms), the parser returned a correct parse tree. This means that the parser itself should be suitable for the task with an application of rule-based amendments in the cases of a few constructions or compound lexemes only.

4. Available NLP-tools and their evaluation on UGC

The task in question was inspired by the challenge of building a sentiment analysis platform (PAKO) for Polish – still quite a resource-poor language in terms of NLP (in spite of the currently biggest WordNet resources worldwide [14]). The National Corpus of the Polish language is an efficient tool, which is still elaborated upon, despite being fairly recent [27]. There are three major morphological analyzers: Morfeusz, Morfologik, and GRAM, which process mostly correct, written-style vocabulary and not able to fully cope with the colloquial Internet discourse. There currently exist two polarity lexica [3, 11], both of which were automatically generated, and neither of which includes the neologisms, colloquialisms, instant messaging language, digits, etc. necessary for opinion-review processing.

The major parsers for Polish (Świgra, Spade/Puddle, and Gobio) when confronted with UGC features, such as diacritics, spelling, and punctuation errors as well as lexica deficiencies, do not provide any results at all (an error message appears in the case of an unknown lexeme in Świgra, which happens in most UGC phrases) or screen up to a hundred results in the case of some difficult sentences (Spejd, which is also frequent). Only Gobio has a useful feature of leaving OOVs beside other parsed phrases and returning all other results of an analysis (POS-tags and phrase types for all other phrases and tokens). Such a solution reduces the parser's performance but makes analysis possible (despite an OOV). All of the above-mentioned parsers are of the constituency type (with HPSG head marking in Gobio), meaning that at the moment, it is hard to find a Polish parser indicating all parts of a sentence, such as predicate, subject, object, attribute, and adverbials. However, these are especially useful in coreference detection (see: [16]) within opinion mining; e.g., in the case of anaphora-based rule building for sentiment extraction expressed within more than one sentence (for example, ***Anna Kowalska*** *used to be a good actress. Now* ***her*** *acting is quite poor*).

Finally, as a Slavic and inflectional language, Polish is often described as difficult in processing due to the free word-order [26], which is why sometimes some of the parsing segmentation rules have to be adjusted (see: [7]).

Of the available morphological analyzers, we have chosen three major tools for the evaluation of UGC-content: Morfeusz[10] (about 4M word forms), Morfologik[11] (about 3,5M word forms or 200,000 lexemes), and GRAM 2.3[12] (about 135,000 lexemes). For the purpose of evaluation, we used the above-described specimen of 55 opinion reviews taken from 4 different social media platforms. As a parameter of evaluation, we used the traditional accuracy measure, defined as the proportion of correctly classified objects to all objects of analysis [21, 22]. We wanted to find out how many lexemes would be assigned a correct POS-tag (true positive). If a word was not found in

[10] `http://sgjp.pl/morfeusz/morfeusz.html`
[11] `http://morfologik.blogspot.com`
[12] `http://www.neurosoft.pl`

a dictionary, we annotated it as a false negative, and when a word was ascribed an incorrect tag, we would treat it as a false positive.

Table 1

Results of accuracy tests of three major Polish morphological analyzers conducted on UGC.

	FOURSQUARE	OPINEO	TRIP ADVISOR	APPSTORE
MORFEUSZ	91.46%	85.46%	86.36%	78.40%
MORFOLOGIK	91.96%	85.29%	84.76%	76.17%
GRAM	87.30%	77.09%	84.28%	70.33%

As you can tell from the above-presented Table 1, there is a difference of performance between the various specimens of opinion reviews. The discourse of Trip Advisor and Foursquare in our specimen are characterized by the least amount of colloquial vocabulary; therefore, they have achieved the highest efficiency scores at POS-tagging and segmentation (not too many unrecognized lexemes, accuracy of 85-92%). The most difficult texts were the opinion reviews from Appstore and Opineo (70-85%), rich in professional as well as colloquial (or even teenager) vocabulary, orthographic mistakes, and lacking in proper punctuation marks.

In the aforementioned test, the best performance ranging from about 78 to 91% in accuracy score can be attributed to Morfeusz, since it has the widest and the most-recently-updated set of vocabulary (last update: 2010). Even though GRAM is the only dictionary recognizing digits and emoticons, the total accuracy of Morfeusz is much higher on all specimens since it contains more colloquial expressions (e.g., *nówka/brand new*). Morfologik performed slightly poorer than Morfeusz; but in general, they share similar vocabulary, and so their results tend to be alike.

For a description of all features, compare Table 2.

Table 2

Differences in functionality features of three major Polish morphological analyzers conducted on UGC.

COMPARATIVE FEATURE	MORFEUSZ	MORFOLOGIK	GRAM
Correction of Polish characters	no	no	no
Correction of spelling mistakes (e.g. JEdyny → jedyny)	yes	no	no
Colloquial lexems (np. *nówka/brand new*, *super/goodish*, *fajne/cool*)	yes	yes	(only some)
Domain-related or professional vocabulary (e.g. hardware)	yes	no	no
Returning all possible tokens	yes	no	no
Lemmatization	yes	yes	yes
Anachronisms	yes	no	no
Emoticons	no	no	yes (basic ones)
Corrector of missing spaces	no	yes	no
Extension of abbreviations (e.g. wg/according to)	yes	yes	yes
Recognition of digits	no	no	yes

A comparison of different features of three major Polish morphological analyzers is of help while determining the best analyzer with respect to the text genre related to UGC. If instant messages are analyzed, recognition of digits and emoticons is inevitable. If colloquial vocabulary prevails (as in opinion reviews), Morfeusz or Morfologik seem to be better choices, as is the case here. Finally, one has to take technical features into consideration. As the content of Morfeusz cannot be amended, it is uncertain whether it would match Gobio as the selected parser. Morfologik can be freely modified, plus it is the default analyzer of the Gobio parser (see below), which meets the demand of the project.

Next, we turned to four major parsers for Polish: the Malt parser[13], Świgra[14], Spejd[15] (otherwise known as Puddle), and Gobio[16]. The first parser is a converter from constituency to dependency output, which requires another parser to conduct analysis and has to undergo training. Although the second tool contains a profound description of all grammatical relations in Polish [35], it returned an "error" message in the case of an OOV. Taking into consideration the above-mentioned UGC characteristics, this means that almost no sentence is parsed because almost all sentences include some challenges. Spejd (Puddle in the PSI-Toolkit version we tested) is a shallow parser very similar to a tagger and sometimes returns about 100 interpretations of a given UGC token. Only the Gobio parser seemed suitable for the task, while providing a proper parse tree for correct-style sentences in a deep-analysis fashion and converting into separate phrase parsing when confronted with an OOV, rough punctuation, and diacritics. Thus, we decided to check whether or not basic improvements would make a difference as to UGC processing by introducing changes to the Gobio parser.

Gobio is a constituency parser with the indication of syntactic heads (Head-driven phrase structure grammar), adjusted with additional algorithms to the free word-order in Polish [34]. It is available for testing on the PSI Toolkit Platform[17], with online access to the parser.

In Table 3, we enlisted results of a sentence and phrase-level segmentation as well as POS-tagging tests conducted on the Gobio parser with the use of the previously described research sample of UGC texts. On the sentence level, we evaluated whether or not the parser recognized the sentence as a whole, which meant that all phrases and tags were correct. On the phrase level, we evaluated both correct segmentation and the proper phrase tag. Correct POS-tags influence the correctness of phrase-tags; therefore, this had to be analyzed as well. Precision is understood as the ratio of correctly tagged sentences or phrases to the number of all phrases or sentences tagged by the parser. Recall is the ratio of the correctly segmented and tagged phrases to the phrases tagged or segmented incorrectly or given no tag. The F1 measure is their harmonic mean [22].

[13] `http://www.maltparser.org`
[14] `http://zil.ipipan.waw.pl/Świgra`
[15] `http://zil.ipipan.waw.pl/Spejd`
[16] `http://psi-toolkit.wmi.amu.edu.pl/help/documentation.html`
[17] `http://www.psitoolkit.pl` [8]

Table 3

Precision, recall and F1 measure results of the Gobio parser with respect to segmentation on sentence and phrase level as well as POS-tagging on a research sample of UGC texts.

	OPINEO	APPSTORE	FOURSQUARE	TRIP ADVISOR
SENTENCE LEVEL				
Precision	12.92%	9.57%	18.18%	13.77%
Recall	33.82%	20.75%	38.89%	38.38%
F1	18.70%	13.10%	24.78%	20.27%
PHRASE LEVEL				
Precision	62.64%	67.66%	78.09%	77.10%
Recall	68.88%	62.78%	78.53%	70.88%
F1	65.61%	65.13%	78.31%	73.86%
POS-TAGGING – ACCURACY	66.13%	58.18%	77.67%	74.22%
SENTENCE + PHRASE LEVEL [21]				
Precision	59.38%	60.10%	73.21%	75.30%
Recall	62.81%	23.08%	71.83%	66.07%
F1	61.05%	33.35%	72.51%	70.38%

We present our results from two perspectives:

- the traditional type of constituency parser evaluation offered by Manning in [21], where the phrase and sentence levels of segmentation are counted altogether,
- with phrase and sentence levels treated separately, since we have noticed the parser's poor performance on the sentence level in case of UGC and good performance on the phrase level.

As we can tell, the difference in precision between the phrase level (60–80%) and sentence level (10–18%) is significant (50–60%). Such a major difference between performance on the sentence and phrase levels is to be attributed to the fact that an unknown lexeme causes the parser to separate the word from the rest of the sentence, whereas a majority of other relations within the sentence remains analyzed correctly. However, the traditional approach in the mode of strict evaluation (discussed by Manning in [21]) does not allow for the perception of differences of performance on various segmentation levels. Based on these results, we assumed that, if we fixed the issue of unknown lexeme recognition (dictionary, orthographic, and spelling deficiencies), we could greatly improve the Gobio parser's performance on UGC.

5. NLP system for Internet opinion parsing in Polish

We approached the problem making a few basic assumptions. First, by manually analyzing a specimen of about 10,000 opinion reviews from social media such as Wizaż, Foursquare, Opineo, TripAdvisor, Appstore, etc., we realized that Internet opinion reviews tend to differ in their character in both Polish and English with respect to the user profile (age and maturity, goal of communication). For example, the majority of Appstore opinions are written by teenagers, where slang and lack of punctuation prevail; whereas Foursquare hotel opinions are provided by quite wealthy, adult busi-

nesspeople, where the lack of Polish diacritics is quite frequent and complex sentence structures occur. We thus assumed that we would not be able to find a common denominator between different opinion sources online unless we conducted a thorough linguistic analysis of a properly selected and balanced sample set of texts.

Secondly, our perception and understanding of what is called *sentiment analysis* differs in comparison with most approaches. Typically, as described by the work by Liu in [18], sentiment analysis refers to positive or negative vocabulary of evaluative character; e.g., *delicious pie* or *unkind restaurant service*, and this is when sentiment lexica can easily be applied to detect adjectives, verbs, or nouns of evaluative character. However, we tend to think in a pragmatic way (as in the Speech Act theory by Searle [33]) that, apart from polarized vocabulary, there are also many syntactic constructions expressing evaluation, sometimes in an implicit way. For example, an order: *Bring back the old version of this app!* if uttered by a customer of an evaluated smartphone application, means he or she is unhappy with its new release (object: *app*, suggestion: *bring back the old version*). Therefore, it could be broader described as *data extraction from opinion reviews* rather than sentiment analysis sensu stricto. To be able to extract such constructions as imperative moods from opinion reviews together with the arguments of the predicate, a deep parser is necessary in order to build data extraction rules on top of the parse tree. What is more, for coreference resolution (e.g., *I bought this app. I think it's really great.*), changes of sentiment (*She used to be a great actress, now she is poor.*) or modal constructions (e.g., *This computer should be better* where the author is implying it is not) deep syntactic analysis is fundamental.

Third, in our experimental approach to build a data-extraction tool, we tried to use currently-available NLP resources for Polish. Out of the four parsers encountered (Świgra, Spejd/Puddle, Malt, and Gobio), we used Gobio because of the following features: 1. it is a deep parser; 2. if unknown lexemes, such as product names, occur in a sentence, it does not present the error message, leaving the sentence unparsed (where Świgra does); 3. its makeup was easy to modify in a rule-based manner. We left its default morphological analyzer – a dictionary with a dehomonimization function – Morfologik, since the results of tests of three available analyzers were similar. None of the analyzers or parsers was able to fully cope with UGC text processing; however, the detection of specific problems made us realize that some changes to specific modules might vastly increase the general performance on opinion reviews. For example, the segmentation rules of the parser were correct but too strict for the punctuation deficiencies of Internet opinion texts.

Fourth, if all of the lexemes were recognized by Gobio, only a few cases (such as *warto/worth*, or compound adverbials such as *na pewno/for sure*) lacked analysis. But when an unknown lexeme such as a proper name appeared, the tree structure was impacted by an unrecognized word. This made us think of the characteristics of the Polish language, such as inflection agreement between; e.g., adjective and noun (same person, case, genus and number) as well as inflection patterns for lexemes with the same suffixes. These features enabled the construction of an inflectional translator to

substitute OOVs with known vocabulary (from the dictionary of Morfologik) to help the parser provide a complete tree. The translation module (inflection provider) spans from name entities (product or company names), through missing regular lexemes (e.g., participles), up to abbreviations.

Such an assumption might seem contrary to frequent text normalization approaches, since we did not perform a typical pre-processing (except for a corrector providing insertion of Polish nouns for digits and amending misspellings). Neither did we implement the machine-learning technique with a corpus normalized for training purposes. This is because a thorough analysis of the Gobio functionalities (e.g., the only parser for Polish enabling unambiguous results of a single opinion sentence) and the parser's performance on opinion texts (sentence parsing even with OOVs) convinced us that it could cope with this task after adaptation.

Having analyzed the available NLP-tools for Polish and the linguistic characteristics of opinion reviews, we estimated the necessary steps to build an opinion-mining system out of these tools.

First of all, we decided to use the **Gobio Segmenter** for pre-segmentation as a default and, thus, compatible tool for the chosen parser. However, in some cases, it was not fitted to the rough punctuation and colloquial sentence structure of UGC. To provide an example, suspension points did not pose a segmentation mark between phrases, and since segmentation marks are frequent in UGC, some sentences were parsed incorrectly; e.g.:

[finite clause with predicate] + [suspension points] + [elliptic clause without subject]
[Podobny model użytkuję od 3 lat][...][Spisuje się bez zarzutu].

Therefore, we built a UGC-segmentation Adapter to deal with these kinds of issues. After detection of such a type of sentence construction, the phrase is split into two sentences with a dot.

For example: *Podobny model użytkuję od 3 lat... Spisuje się bez zarzutu.*

→ *Podobny model użytkuję od 3 lat...*
→ *Spisuje się bez zarzutu.*

The same level of analysis deals with incorrect punctuation problems. In the cases where typically no parse tree would be built due to wrong punctuation (e.g., the lack of a capital letter), segmentation rules were adjusted to the cases of possible sentence splitting after a punctuation mark followed by a lowercase letter. Emoticons were defined as separate entities with attributed meanings instead of an aggregation of punctuation marks subject to separate analysis. At the same stage of processing, some of the hindrances to Gobio's tree building had to be removed, such as markers of text and sentence coherence; e.g., *Ponadto/What is more*, *W dodatku/On top of that*, and other linking words yet unrecognized by the parser but unnecessary for analysis.

Secondly, the use of a corrector (**ikorektor**) was inevitable in order to bring back the Polish characters to texts written without them. Since Gobio was made

for processing the stylistically and grammatically correct texts (trained on correct style corpora), it is impossible to analyze colloquial language texts without specific characters. However, on top of that, we also had to build a UGC-spelling Adapter, since the correction tool sometimes added Polish characters where it was unnecessary. Therefore, part of the ikorektor rules had to be amended; e.g., *pozytywny* (positive) instead of *pożytywny* (an English equivalent would be some type of a misspelling of the word *positive*, such as *possitive*).

Then, we picked one of the above-described morphological analyzers – **Morfologik**, compatible with the parser that obtained the highest test score – Gobio, and prepared a rule-based Inflection Provider for unrecognized words. Colloquialisms or other unknown vocabulary were replaced with synonyms of a similar inflection pattern of the same morphological class; e.g., the unrecognized *troszkę* (tiny little bit) is substituted with the known *trochę* (a little bit) for parsing purposes and then brought back to its original form[18]. Since no NER-databases containing proper names of products were available for Polish, we have used this inflectional translator also to exchange the unknown proper names with the recognized equivalents of their categories, having full inflectional patterns.

For example:

W Butchery&Wine mają świetne wino./Butchery&Wine has great wine.
→ W miejscu [Butchery&Wine] mają świetne wino./
Place [Butchery&Wine] has great wine.

The Inflection Provider works on two levels of analysis – first, after the correction of spelling, an unrecognized brand name is replaced by an analyzable word; and second, after a correct parse tree is built, the brand name is returned to the sentence in question. This method seems suitable, especially when dictionary amendments are time-consuming or impossible due to the closed character of its library.

An interesting method of providing product categories we are currently considering is their online crawling on popular opinion portals to cater for unknown labels (e.g., Opineo: HP nx 2700 – category laptop). The method is similar to obtaining aspects and features of products online, which was offered e.g., by Hu and Liu in [12]. Nevertheless, it is different in its implementation, since the OOVs in the text are exchanged with their categories, typically common words that can be easily found in any morphological analyzer. In this way, an OOV becomes recognized, and additional semantic knowledge of a category is added to the system. This means that it is possible to build a NER-tool by crawling the available product categories from the Web (Web crawler) [17].

[18]We are currently in the process of amending this step of processing with respect to some of the enlisted vocabulary omissions, especially digits, emoticons, and IM-language. Some of the already-included words still have to be corrected in terms of their category (e.g., *który/which* memorized within Morfeusz and Morfologik as adjective deserves the proper category of a pronoun). GRAM, on the other hand, would require an amendment of a category of linking words and particles (instead of a general category "other").

Apart from the adjustment of the parser and analyzer to the UGC-content, other adjustments had to be made to enable sentiment analysis:

- The system is based on an adaptation of a constituency parser to the Polish language. Not only is it a dictionary of the morphological analyzer (Morfologik) that had to be amended based on the above-mentioned lexical resources, but also some of the parser's grammar rules. One of the issues in question were participles, as they are not included in the grammatical system. Nevertheless, in a simplified grammar, some of them can be treated as adjectives and parsed.
- Additionally, a list of homonyms derived from incorrect spelling had to be identified, e.g. (*tez/theses-też/also*; *nie ma/no* or *none-niema/deaf*), as well as some words with additional definitions for their new colloquial uses; e.g., *je* (either colloquial *is* or *is eating*).
- Finally, we had to create a sentiment dictionary dedicated to the domains of accommodation and dining, since the available digital sentiment dictionary of Polish covered other domains and would bring in a different sentiment scale than expected.

Figure 1 shows three major layers of the system. A text of an opinion review (crawled from social media) enters the parser to be directed to the first level of segmentation, correction, and inflection-fitting. This preparatory stage enables its adjustment to morphological analysis in Morfologik and disambiguation and parsing by Gobio. A parse tree is built, and its output is further converted into a moderately language-independent simplified phrase format that enters a model of sentiment analysis (to be presented in future works). The amended parser architecture is presented in detail in Figure 2.

6. Experimental studies

In order to check if our solutions have brought an increase in the efficiency of the parser, we prepared a new specimen of texts for testing. It was of the same makeup as the training one; meaning: the same Internet genre (opinion reviews), the same or similar online resources (Opineo, Tripadvisor, Foursquare, Appstore), and text proportions. We aimed for the following goals:

- to measure the increase in POS-tagging due to the implementation of the inflection provider and correction module (with the accuracy measure),
- to find out how our pre-processing and parser adaptation solutions have impacted the parser's performance on the phrase and sentence levels (with the same parameters of precision, recall, F1 on different set samples, and altogether),
- to determine the increase of the number of correct parse trees among all trees built (with the accuracy measure) in order to detect the types of parser proper errors, independent of text type and resulting from unknown grammar constructions remaining after pre-processing and improved parsing.

As we can tell from the results presented in Table 4 and along Manning's line of argument (see: [21]), the hypothesis has been confirmed that, if the major type

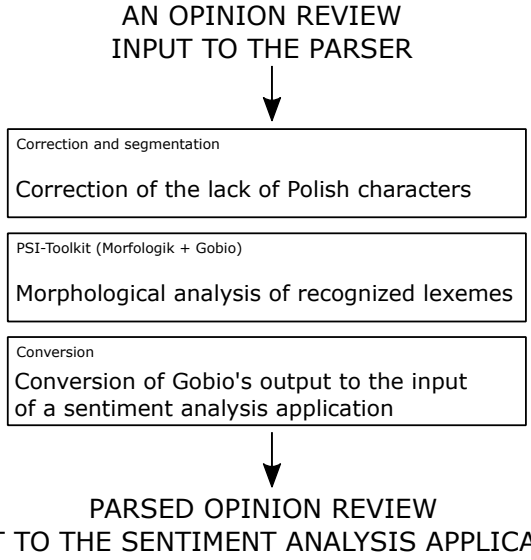

Figure 1. Major levels of the Gobio parser adapted to the processing of opinion reviews in Polish.

Table 4

Test results of the Gobio parser after text normalization and parser adaptation to UGC conducted on a parallel specimen of opinion texts.

	OPINEO	APPSTORE	FOURSQUARE	TRIP ADVISOR	ALL TYPES TOGETHER
POS-ACCURACY	0.97	0.99	0.98	0.98	0.98
PRECISION	0.94	0.88	0.85	0.93	0.89
RECALL	0.96	0.91	0.90	0.95	0.93
F1	0.95	0.89	0.87	0.94	0.91

of error (here: lexical) is amended, the general parser performance increases. This is visible both in the POS-tagging accuracy and in the F1 measure since, if almost all tokens are recognized (general POS-tagging accuracy of 0.98), the general parser performance (F1 measure of 0.91) develops. In the Appstore and Foursquare specimens, there was another factor influencing the parser's performance; namely, very relaxed segmentation and punctuation that was difficult to deal with, which is why the precision and F1 scores are slightly worse. However, in general, it seems that a profound linguistic analysis of UGC-specific phenomena (problematic for the parser) and proper solutions designed to reverse their negative effect bring significant improvements to the overall score.

On the other hand, despite high POS-accuracy as well as the provision of Polish diacritics and relaxed segmentation, some problems still remain (which probably are related to the parser itself). In order to detect such difficult syntactic constructions,

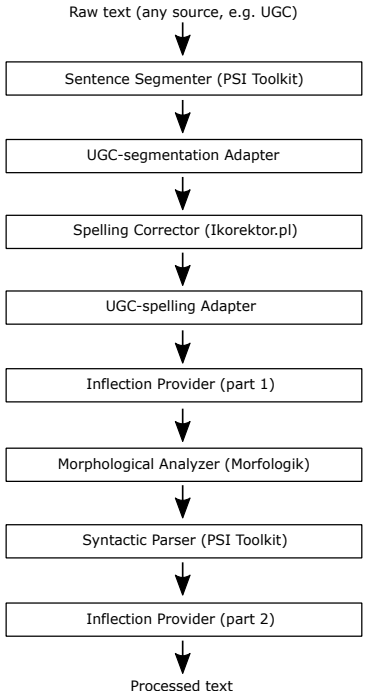

Figure 2. Architecture of an NLP system for Internet opinion parsing in Polish.

we carried out a test. This time, we wanted to check the number of correctly parsed trees; by *correct* we mean:

- the entire tree built correctly in terms of segmentation on both phrase and sentence levels,
- without omissions of unrecognized vocabulary, misspellings, etc.,
- with correct Polish characters included.

In other words, we wanted to find out how the implemented modules of analysis impacted the correctness of the whole tree structure and how the error propagation effect could be reversed (see: [21]). As we can tell from Table 5, the solutions offered above have brought a major change in the efficiency of the presented parser.

The accuracy measure used (defined as number of correctly built trees divided by the number of trees built correctly and incorrectly) provides us with information that the correct parsing of UGC has increased significantly (double in the case of three text sources, and triple in the case of one specimen) together with the recognition of the previously unknown vocabulary and introduction of correct punctuation, segmentation, and spelling correction, as well as inflection translation.

Despite the implementation of all of these improvements and high POS-tagging results, the reason why about one fifth of the sentences causes difficulties in the par-

Table 5

Test results of the Gobio parser before and after text normalization and parser adaptation to UGC.

	OPINEO	APPSTORE	FOURSQUARE	TRIP ADVISOR
Ratio of the number of entire correct trees to all trees built before pre-processing and parser adaptation	38.89%	30.43%	17.54%	36.36%
Ratio of entire correct trees to all trees built before pre-processing and parser adaptation	77.78%	78.26%	63.16%	76.36%

ser's building of the entire parse tree can be inferred from problematic cases. One of them could be – as indicated above – that certain Gobio segmentation or parsing rules have to be amended for compound-sentence analysis. Many times, compound sentences (e.g., with *który/which*) tend to be split instead of parsed together. Also, compound sentences with verbless phrases or less-typical predicates [e.g., *warto*, *oby* (English equivalents of *worth* and *may*) + *infinitive*] are not recognized by the parser. Another reason might be that not all vocabulary combinations were present in the training corpus, which is why sometimes the same grammatical rules do not apply to all verbs (e.g., *Lubię/I like + object* gives a proper parse tree; however, *Polecam/I recommend + object* often splits sentence segmentation). Some lacks in basic vocabulary should also be amended, like e.g., *wiele/wielu* (English for *many*) or compound adjectives (e.g., *cytrynowo-śmietankowy* – English for *lemon-cream*). Sometimes, an atypical word order causes the parser to fail to build a complete tree. A broader re-training corpus or further rule-based amendments could be used to make up for these minor features.

7. Conclusions

In this work, we have proven that, in resource-deficient languages, a thorough investigation of linguistic features of opinion reviews and the appropriate design of their adjustment in a rule-based fashion can lead to their application in the UGC domain without the necessity of preparing new, normalized corpora for parser training. Solutions built on three levels (correction, segmentation, inflection-based translation) with recognition of Polish grammar result in a significant improvement of parsing results of UGC.

To the best of our knowledge, this piece of work poses one of the first attempts of adjusting the available Polish NLP tools to UGC-processing in opinion mining. Its insights seem important with respect to thorough text analysis conveyed, linguistic phenomena identified and described, as well as an efficient rule-based solution offered. In the case of languages with scarce NLP-tools and growing and impatient market demand (such as Polish), development and efficient adjustment of existing systems is valued. Finally, in a broader perspective, such an implementation enables the fitting

of new languages to a fairly language-independent system of data extraction from texts via implementation and adaptation of their parsers.

As far as future works are concerned, further domains of UGC could also be looked into, and the above-described NLP system applied to sentiment analysis application designed within the PAKO platform[19].

References

[1] Aw A., Zhang M., Xiao J., Su J.: A phrase-based statistical model for SMS text normalization. In: *Proceedings of the COLING/ACL on Main conference poster sessions*, pp. 33–40, Association for Computational Linguistics, 2006.

[2] Beaufort R., Roekhaut S., Cougnon L.A., Fairon C.: A hybrid rule/model-based finite-state framework for normalizing SMS messages. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 770–779, Association for Computational Linguistics, 2010.

[3] Buczynski A., Wawer A.: Shallow parsing in sentiment analysis of product reviews. In: *Proceedings of the Partial Parsing workshop at LREC*, vol. 2008, pp. 14–18, 2008.

[4] Chiticariu L., Li Y., Reiss F.R.: Rule-Based Information Extraction is Dead! Long Live Rule-Based Information Extraction Systems! In: *EMNLP*, pp. 827–832, 2013.

[5] Choudhury M., Saraf R., Jain V., Mukherjee A., Sarkar S., Basu A.: Investigation and modeling of the structure of texting language. *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 10(3–4), pp. 157–174, 2007.

[6] Cook P., Stevenson S.: An unsupervised model for text message normalization. In: *Proceedings of the workshop on computational approaches to linguistic creativity*, pp. 71–78, Association for Computational Linguistics, 2009.

[7] Graliński F.: *Formalizacja nieciągłości zdań przy zastosowaniu rozszerzonej gramatyki bezkontekstowej*. Ph.D. thesis, Adam Mickiewicz University, Faculty of Mathematics and Computer Science, Poznań, 2007.

[8] Graliński F., Jassem K., Junczys-Dowmunt M.: PSI-toolkit: A natural language processing pipeline. In: A. Przepiórkowski, M. Piasecki, K. Jassem, P. Fuglewicz, eds., *Computational Linguistics*, *Studies in Computational Intelligence*, vol. 458, pp. 27–39, Springer, 2013.

[9] Grzenia J.: *Komunikacja językowa w Internecie*. Wydawnictwo Naukowe PWN, Warszawa, 2006.

[10] Gunelius S.: The data explosion in 2014 minute by minute – Infographic. *Newstex*, vol. 12(07), 2014.

[11] Haniewicz K., Kaczmarek M., Adamczyk M., Rutkowski W.: Polarity lexicon for the polish language: Design and extension with random walk algorithm. In: *Advances in Systems Science*, pp. 173–182, Springer, 2014.

[19]Grant no INNOTECH-K2/IN2/89/182461/NCBR/13 of the National Center of Research and Development.

[12] Hu M., Liu B.: Mining opinion features in customer reviews. In: *AAAI*, vol. 4, pp. 755–760, 2004.

[13] Hwa R.: Sample selection for statistical parsing. *Computational Linguistics*, vol. 30(3), pp. 253–276, 2004.

[14] Kędzia P., Piasecki M., Orlińska M.: Word Sense Disambiguation Based on Large Scale Polish CLARIN Heterogeneous Lexical Resources. *Cognitive Studies*, (15), pp. 269–292, 2015, http://dx.doi.org/10.11649/cs.2015.019.

[15] Kobus C., Yvon F., Damnati G.: Normalizing SMS: are two metaphors better than one? In: *Proceedings of the 22nd International Conference on Computational Linguistics*, vol. 1, pp. 441–448, Association for Computational Linguistics, 2008.

[16] Kopeć M.: Polski Korpus Koreferencyjny – wersja 0.85. 2013, http://zil.ipipan.waw.pl/PolishCoreferenceCorpus.

[17] Krupa T.: Studium przypadku – system ISPAD. In: B. Wiszniewski, ed., *Inteligentne wydobywanie informacji ze społecznościowych serwisów internetowych*, Automatyka i Informatyka. Technologie Informacyjne. Internet i Sieci Semantyczne, pp. 121–139, PWNT, 2011.

[18] Liu B.: Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, vol. 5(1), pp. 1–167, 2012.

[19] Luo W., Litman D.J., Chan J.: Reducing Annotation Effort on Unbalanced Corpus based on Cost Matrix. In: *HLT-NAACL*, pp. 8–15, 2013.

[20] Manning C.: Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In: *Computational Linguistics and Intelligent Text Processing*, pp. 171–189, Springer, 2011.

[21] Manning C.: Evaluation of Constituency Parsers. Stanford lectures online. 2012, http://www.youtube.com/watch?v=mMXgbrts82M.

[22] Manning C., Schütze H.: *Foundations of statistical natural language processing*. MIT press, 1999.

[23] Martínez P., Segura I., Declerck T., Martínez J.L.: TrendMiner: Large-scale Cross-lingual Trend Mining Summarization of Real-time Media Streams. *Procesamiento del Lenguaje Natural*, vol. 53, pp. 163–166, 2014.

[24] Nagarajan M., Gamon M.: Workshop on Language and Social Media – Introduction. In: *Proceedings of LSM 2011*, 2011.

[25] Ong W.J.: *Orality and literacy*. Routledge, 2013.

[26] Piasecki M.: Polish tagger TaKIPI: Rule based construction and optimisation. *Task Quarterly*, vol. 11(1–2), pp. 151–167, 2007.

[27] Przepiórkowski A., Bańko M., Górski R.L., Lewandowska-Tomaszczyk B.: Narodowy Korpus Języka Polskiego. 2012, www.nkjp.pl.

[28] Raghunathan K., Lee H., Rangarajan S., Chambers N., Surdeanu M., Jurafsky D., Manning C.: A multi-pass sieve for coreference resolution. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 492–501, Association for Computational Linguistics, 2010.

[29] Ratuszniak B.: Monitoring social media. Co oferują firmy? [online], 2012, http://goo.gl/8p7mGp, accessed: 25.04.2012.

[30] Ray A.: Customer Affinity Meets Brand Vectors: Sentiment that Matters. 2013, sentiment Analysis Symposium, New York.

[31] Read J., Flickinger D., Dridan R., Oepen S., Øvrelid L.: The WeSearch Corpus, Treebank, and Treecache. A comprehensive sample of user-generated content. In: *In Proceedings of the 8th International Conference on Language Resources and Evaluation*, Citeseer, 2012.

[32] Rodrıguez-Penagos C., Atserias J., Codina-Filba J., Garcıa-Narbona D., Grivolla J., Lambert P., Saurı R.: Combining lexicon-based ML and heuristics. In: *Second Joint Conference on Lexical and Computational Semantics (*SEM)*, vol. 2, pp. 483–489, Association for Computational Linguistics, 2013.

[33] Searle J.R.: *Speech acts: An essay in the philosophy of language*, vol. 626. Cambridge University Press, 1969.

[34] Skórzewski P.: Gobio and PSI-Toolkit: Adapting a deep parser to an NLP toolkit. In: Z. Vetulani, H. Uszkoreit, eds., *Proceedings of the 6th Language and Technology Conference*, pp. 523–526, Fundacja UAM, Poznań, 2013.

[35] Świdziński M.: *Gramatyka formalna języka polskiego*. Wydawnictwo Uniwersytetu Warszawskiego, 1992.

[36] Van Hee C., Van de Kauter M., De Clercq O., Lefever E., Hoste V.: LT3: Sentiment Classification in User-Generated Content Using a Rich Feature Set. In: *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 406–410, Association for Computational Linguistics and Dublin City University, Dublin, Ireland, 2014.

[37] Zdunkiewicz-Jedynak D., Ciunovič M.: *Ćwiczenia ze stylistyki*. Wydawnictwo Naukowe PWN, 2010.

[38] Zhenzhen X., Dawei Y., Brian D.D.: Normalizing microtext. In: *Proceedings of the AAAI-11 Workshop on Analyzing Microtext. San Francisco, AAAI*, pp. 74–79, 2011.

Affiliations

Agnieszka Pluwak
Institute of Slavic Studies, Polish Academy of Sciences, Warsaw, Poland
Fido Intelligence, Gdansk, Poland, agnieszka.pluwak@gmail.com

Wojciech Korczynski
AGH University of Science and Technology, Faculty of Computer Science, Electronics and Telecommunications, Department of Computer Science, Krakow, Poland,
wojciech.korczynski@agh.edu.pl

Marek Kisiel-Dorohinicki
AGH University of Science and Technology, Faculty of Computer Science, Electronics and Telecommunications, Department of Computer Science, Krakow, Poland, doroh@agh.edu.pl

Received: 28.11.2014
Revised: 12.05.2015
Accepted: 16.09.2015