Inforex — a Collaborative System for Text Corpora Annotation and Analysis

Michał Marcińczuk **Marcin Oleksy** **Jan Kocoń**
G4.19 Research Group
Department of Computational Intelligence
Faculty of Computer Science and Management
Wrocław University of Technology, Wrocław, Poland
{michal.marcinczuk,marcin.oleksy,jan.kocon}@pwr.edu.pl

Abstract

We report a first major upgrade of Inforex — a web-based system for qualitative and collaborative text corpora annotation and analysis. Inforex is a part of Polish CLARIN infrastructure[1]. It is integrated with a digital repository for storing and publishing language resources[2] and it allows to visualize, browse and annotate text corpora stored in the repository. As a result of a series of workshops for researchers in Humanities and Social Sciences we improved the graphical interface to make the system more friendly and readable for non-experienced users. We also implemented a new functionality for a gold standard annotation which includes private annotations and annotation agreement by a super-annotator.

1 Introduction

Digital humanities (DH) create new demand and challenges for development of new or existing tools and systems for text documents manipulation, processing, analysis and visualization. CLARIN-PL — the Polish part of CLARIN infrastructure — tries to rise the challenges associated with DH for Polish language. Among many other issues, there is a need for an intuitive and easy to use system for qualitative text corpora management, annotation, analysis and visualization. To fulfill these needs we develop such a system called Inforex. In this article we present the current state of the system development.

The decision to create a system for text corpora annotation was taken in 2009 when there were no such systems which support collaborative work. On that time the only existing tools were desktop applications for individual work such as GATE (Cunningham et al., 2011) or Manufakturzysta Luna (Marciniak et al., 2010). Since 2010 several systems have emerged, like WebAnno 3 (Eckart de Castilho et al., 2016) or GATE Teamware (Bontcheva et al., 2013).

The first version of Inforex system was released in 2010 and its initial role was to construct corpus-based linguistic resource for various tasks from the field of natural language processing, including named entity recognition (Marcińczuk et al., 2011), shallow parsing (Radziszewski and Piasecki, 2010), word sense disambiguation (Bas et al., 2008), recognition of semantic relations between named entities (Marcińczuk and Ptak, 2012). It was used to develop two major (at that time) resources for Polish: Corpus of Wrocław University of Technology called KWPr (Broda et al., 2012) (within the NEKST[3] project) and Corpus of Economic News (CEN) (Marcińczuk et al., 2013) (within the SyNaT project[4]). Later, in 2013 Inforex was used to construct another major resource, which is Polish Corpus of Suicide Notes (PCSN)[5] (Marcińczuk et al., 2011) guided by Monika Zaśko-Zielińska (2013). Until now the system has been used to access the corpus. The access is granted on a demand after obtaining a permission form Wrocław University.

In 2013 Poland joined CLARIN — European Research Infrastructure for Language Resources and Technology. The goal of CLARIN is to make the language technologies more accessible to researches from humanities and social sciences, which in most cases do not have the technical skills to use many of the tools on their own. At that time we made a decision to make Inforex a part

[1] http://clarin-pl.eu
[2] http://clarin-pl.eu/dspace
[3] http://nekst.ipipan.waw.pl/
[4] http://www.synat.pl/
[5] http://pcsn.uni.wroc.pl/

Proceedings of Recent Advances in Natural Language Processing, pages 473–482,
Varna, Bulgaria, Sep 4–6 2017.
https://doi.org/10.26615/978-954-452-049-6_063

of the Polish CLARIN infrastructure. In 2015–2017 we have organized several workshops for researchers in humanities and social sciences. The workshops showed us several user experience issues. System GUI turned out to be not enough intuitive for non-experienced users. Then, first of all, it needed to be simplified. Second problem was connected with the methodology. The researchers use various tools for corpora analysis (including spreadsheets) and Inforex may be treated as some kind of pre-processing tool that allows to prepare corpus for further analysis. Data export was possible but complicated and required an access to a database. Users feedback proved that the easy form of data export is one of the crucial needs. After the set of workshops we gathered more information about other important needs (also in the form of questionnaires) like access to a custom annotation schemas definition or data visualisation. Some of them have been already implemented and the other are under construction.

2 Inforex Features Overview

In the following sections we present the main functionalities and features of the Inforex system.

2.1 Web-based Access

Inforex is a web-based tool which does not require installation. It can be accessed by any web-browser which support JavaScript. Despite Inforex is built on several universal JavaScript libraries and frameworks (jQuery, jQuery extensions and Bootstrap) we suggest using Chrome and Firefox. These two web browsers are used to test the system on daily bases. Users might use other browsers as well, however we are not able to validate all functions in each of the available web browsers, thus some minor issues might occur.

2.2 Authorized and Public Access

Corpora stored in Inforex can be accessed by authorized and unauthorized users. The manager of the corpus (the owner or a user with specific privileges) decides what type of information from the corpora can be publicly available. For instance, only authorized users can have access to documents' content and can modify the corpus annotations while unauthorized users may have access to some statistics or annotation frequency lists.

2.3 Integration with DSpace as a Part of Polish CLARIN Infrastructure

Inforex system is available at http://inforex.clarin-pl.eu and it is part of Polish CLARIN infrastructure. This installation is integrated with the official repository for language resources in Polish CLARIN[6]. The repository runs on DSpace system[7]. When a user registers in https://clarin-pl.eu/dspace/, he also gains access to Inforex system. At this stage accounts are automatically synchronized. In the future both systems will use unified federation authorization.

2.4 Collaboration

Inforex offers several ways for collaborative work on a single corpus. One of them is the access to the same corpora for different authorized users. The other one is a selective, task-oriented access to the same document. For instance, different groups of users can have access to document's metadata. The last one is the "2+1" annotation, i.e. two or more users annotate the same set of documents independently and the super-annotator creates the final set of annotations based on their input. More about this type of collaboration is presented in Section 3.2.

2.5 Qualitative Document Annotation

Inforex was designed for qualitative document annotation. This means it does not offer a fast and robust search functions over large corpora with millions of documents. Such functionality can be obtained using other existing tools designed for it, for instance Sketch Engine (Kilgarriff et al., 2014) or NoSketch Engine (Rychlý, 2007). Inforex is suited for medium size corpora (containing thousands of small documents) and to manually describe documents in terms of their metadata, annotations (types of phrases organized in a hierarchy), annotation attributes, relations between annotations and annotation frames.

2.6 Language-independent

Inforex is language-independent in the sense that it can handle documents in any natural language. So far it has been used to annotate Polish, English and Hebrew texts (see Section 3.2).

[6] https://clarin-pl.eu/dspace/
[7] https://github.com/ufal/clarin-dspace

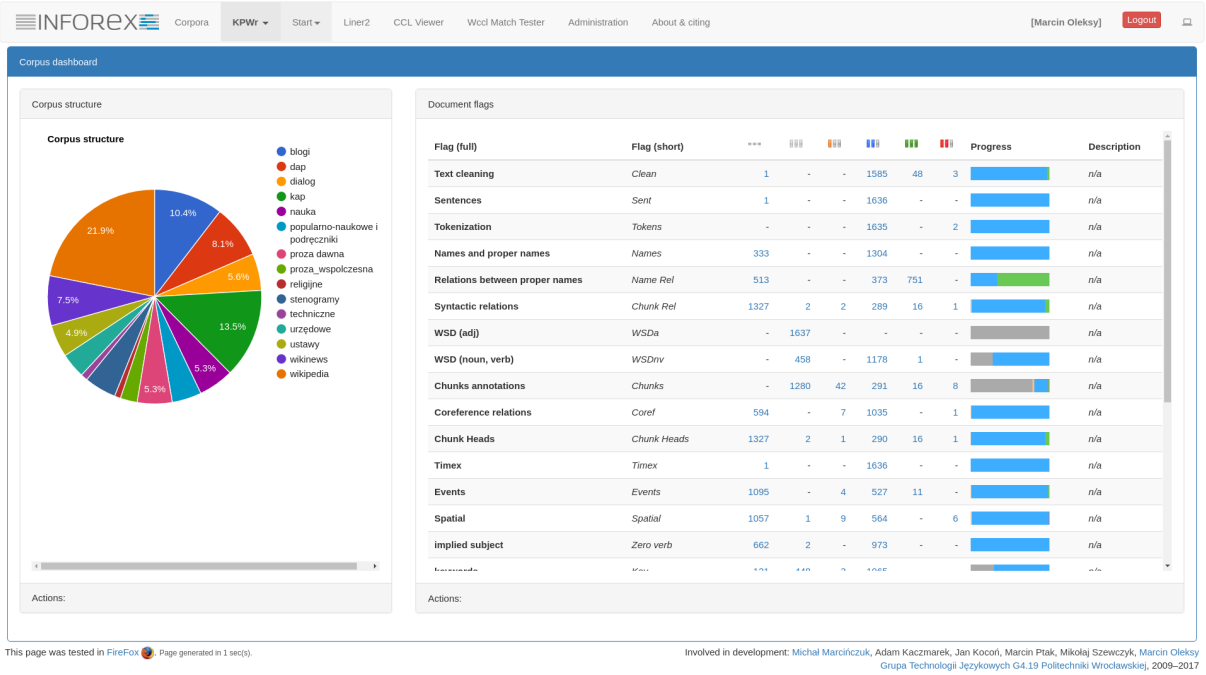


Figure 1: Corpus overview

2.7 Document Visualisation

Inforex can handle documents in two formats: plain text and XML. For XML documents it is possible to display their content in a visually formated way. This allows to highlight the document structure what improves the user experience while browsing and annotating documents. Sample visualizations of different types of documents are presented in Figure 3.

2.8 Document Description

Inforex supports four types of information units which can be used to describe documents content:

1. Metadata — an information unit which is assigned to whole document (author name, document creation time, source, etc.).

2. Annotation — an information unit which is assigned to a sequence of words in the document content. Each annotation is described with a category (categories can be organized in a hierarchy) and a set of attributes. The set of attributes depends on the semantic interpretation of the annotation category. For instance, for named entities it can be a lemma, for temporal expressions it can be a normalized value of the expression and for event mentions it can be an event modality.

3. Relation — an information unit which is assigned to a pair of annotations. It is a directed link between two annotations of some category.

4. Frame — an information unit which is assigned to a set of annotations. Frame consists of a set of annotations with roles assigned to them. This type of structure can be used for event annotations (LCD, 2005).

3 Recent Improvements

In the following sections we present the recent major improvements of Inforex system.

3.1 Modern Layout

A set of workshops carried out from 2015 to 2017 showed that there was the need for an adjustment of user interface to a new group of users — researchers in humanities and social sciences not involved in NLP tools development. New users reported confusion with the large amount of information and the number of available functions. The need of interface simplification appeared while functionalities of the system would remain unchanged. Thus, Inforex layout has been upgraded and modernized. It involved not only a design lifting of the user interface but also changes in navigation panels. The comparision of *old* and *new*

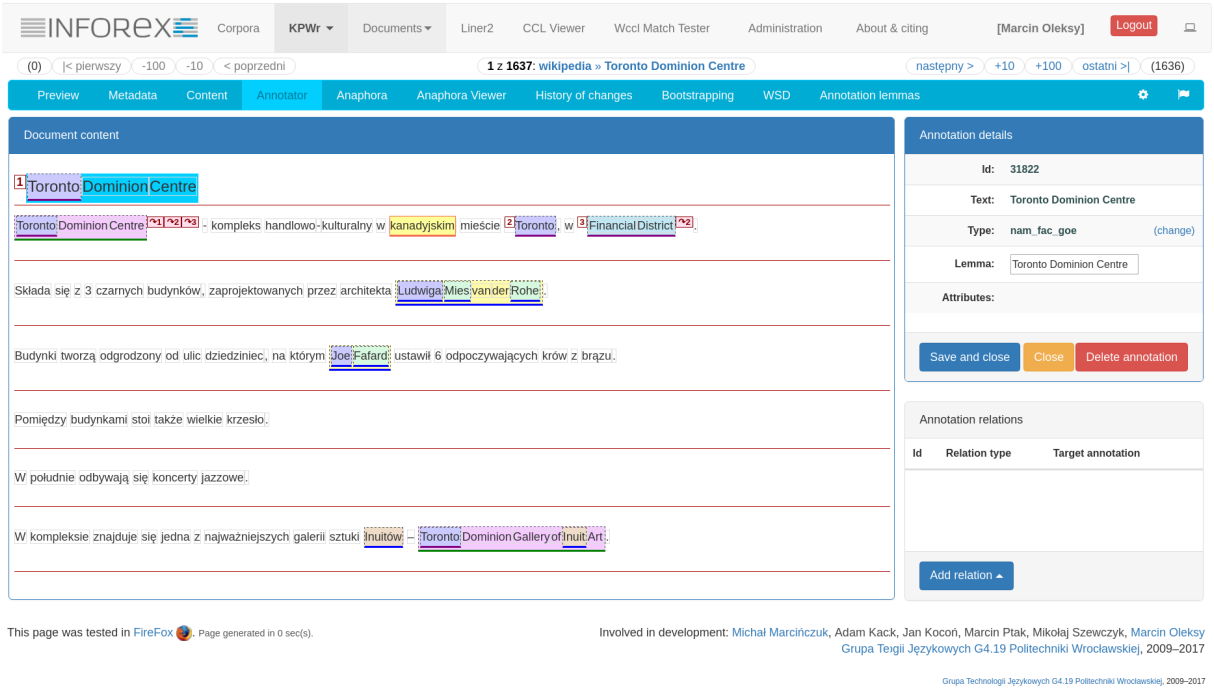


Figure 2: Document annotation view

layout is presented in Figure 4.

3.2 Annotation Agreement

Reliability is a key value in the creation of a good quality corpora for learning and testing of NLP tools. The current version of Inforex enables simultaneous and independent annotation of the same text sample by more than one annotator. Moreover, the annotation process coordinator may keep track of inter-annotator agreement between two raters thanks to the *Agreement module* which uses Positive Specific Agreement (PSA) measure (Hripcsak and Rothschild, 2005) to calculate the reliability (see Figure 5). View configuration gives the opportunity to define annotation layers, subsets or categories, users and set of documents that have to be analysed. The coordinator may also specify a comparison mode: whether the system has to take into consideration the annotation boundaries only or boundaries and categories. It may also include annotation lemmas. Inter-annotator agreement is a very important indicator of the annotation guidelines clearness or cohesion. Keeping track of changes of the inter-annotator agreement between subsequent annotation iterations helps to improve the quality of the annotation guidelines. Agreement module makes that process easier and faster.

Inforex system also supports the curation of the annotation process (see Figure 6). The curator can make choice between two different annotators choices, or even reject consistent but incorrect annotations. Thanks to that module several Gold Standard projects were performed e.g. Polish Coreference Corpus (Ogrodniczuk et al., 2015) for definite descriptions annotation and Polish Spatial Texts corpus for the annotation of dynamic spatial expressions.

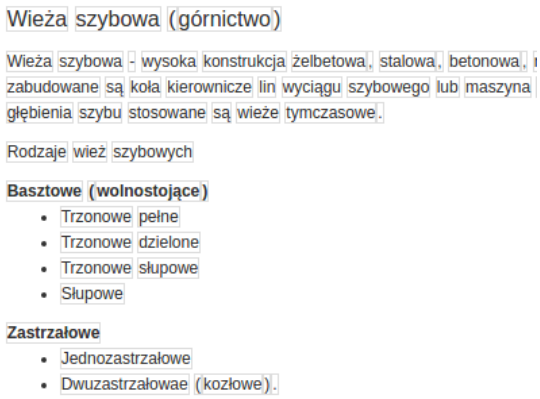
4 Applications

In the following sections we present several practical applications of the Inforex system.

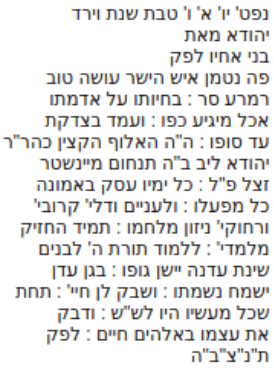
4.1 KPWr

KPWr (Polish Corpus of Wrocław University of Technology) (Broda et al., 2012) is a corpus of written and spoken documents available on the Creative Commons license which is intended primarily as a training and testing material for NLP tools being developed at Wrocław University of Science and Technology. It is successively enriched with annotation layers. Inforex recently supported manual text annotation within such layers as temporal expressions and their normalizations, events (and description of event attributes), spatial expressions and semantic roles. In order to prepare temporal expressions annotation (Kocoń et al., 2015) a new annotation scheme based on

(a) Facebook conversation.



(b) Wikipedia article.



(c) Hebrew document.

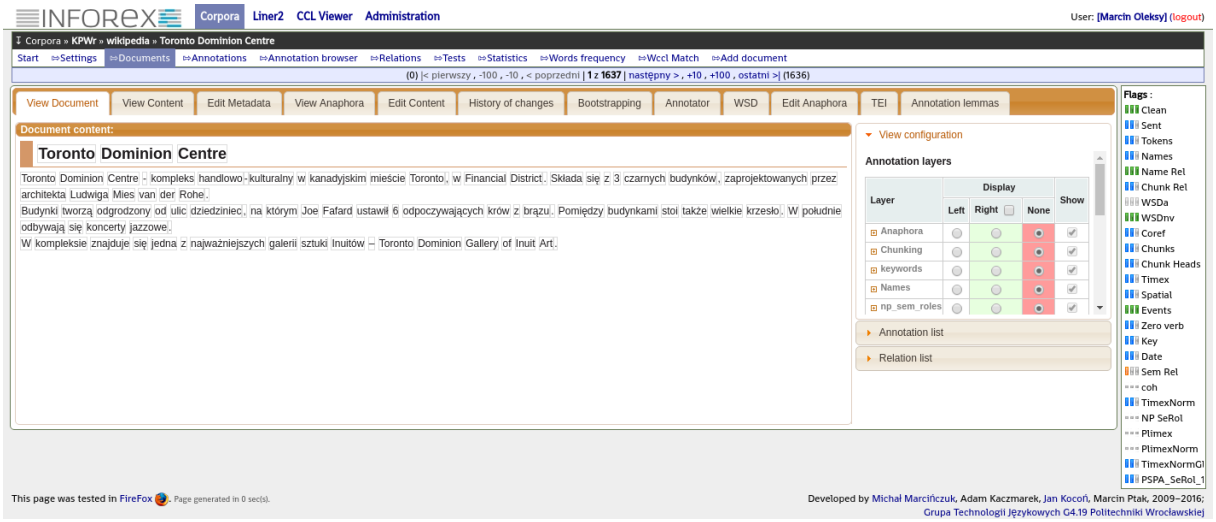
Figure 3: Sample documents visualizations

TimeML was added. These categories refer to a date, time of a day, duration and frequency of an event. Annotation lemmas perspective was used to provide normalized temporal expressions, revealing that the term 'lemma' in Inforex may function as a broad concept. The Annotator perspective from the system also supports event annotation (Marcińczuk et al., 2015). There are seven coarse-grained categories of events, i.e. action, state, reporting, perception, aspectual, intensional action and intensional state. The categorization was based on the TimeML guidelines with some modifications. It also involved creation of a new annotation scheme. The flexibility in adding new annotation layers (setting the new annotation categories) is one of the most important features. The possibility of establishing relations between annotated fragments is not less relevant. It was crucial e.g. for spatial expressions annotation. Its main goal was to extract different ways of distributing spatial information throughout a sentence by reviewing the lexical and grammatical signals of various relations between objects (Marcińczuk et al., 2016).

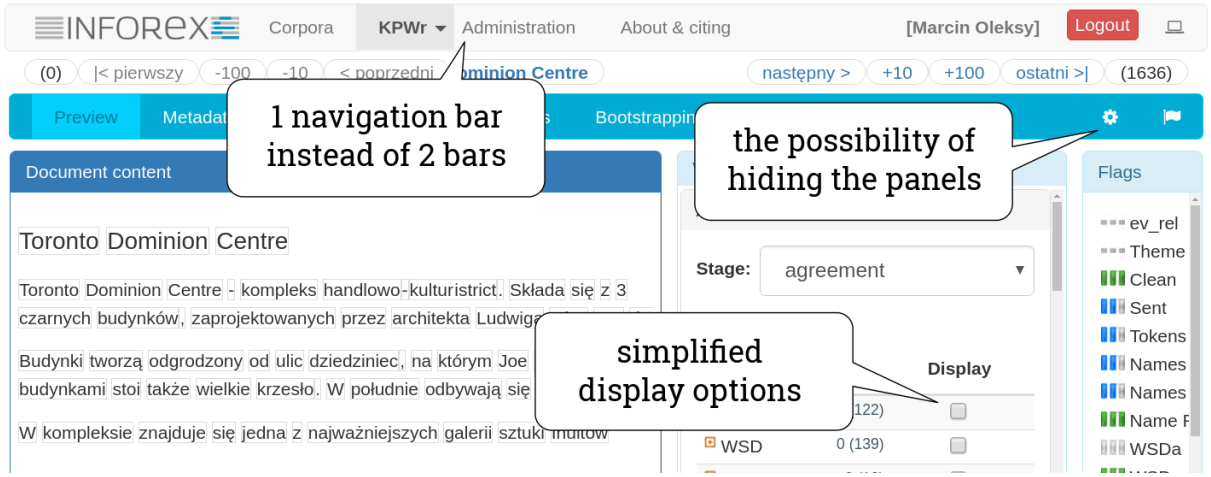
4.2 European Legal Texts

As practice shows, although Inforex was primarily developed for Polish language, that it can also be used to work with documents written in other languages. Inforex features and functionalities are useful e.g. in examining current EU official literature related to territorial development and urban planning. Authors of this analysis first uploaded EU Territorial Policy Documents 2007-2016[8] to CLARIN-PL DSpace repository and then imported it to the Inforex system. The corpus was divided into 4 subcorpora and prepared for qualitative and quantitative analysis. The review of the key strands enabled the identification of its 8 core values (or principles) for further statistical and contextual analysis. After ascribing to each category its textual triggers (word forms), a quantitive analysis using words frequency lists generated by Inforex was performed. Manual annotation with a newly defined set of annotations and Annotation Browser with the possibility of exporting data were a great support for qualitative analysis — detailed contextual analysis of the corpus focused on two crucial categories: *Participation* and *Communication*.

[8] http://hdl.handle.net/11321/316



(a) Inforex layout before modernization



(b) Inforex layout after modernization

Figure 4: Inforex layouts comparison

4.3 Hebrew Corpus

Inforex supports manual annotation even if the text is written using non-latin alphabet and a right-to-left notation. One of the system applications was related to a corpus of Hebrew gravestone inscriptions. It also involved the creation of a new annotation schema. Categories referred mainly to the pragmatic level of communication (e.g. initial and final expressions, laudations, death circumstances). The perspective of annotation lemmas was used to enter Polish translations of annotated fragments, which also showed that the lemma attribute may be a broad term especially in the case of practical applications of the system.

4.4 Other Corpora

Inforex was used to prepare the training data during participation in BSNLP 2017 shared task on multilingual named entity recognition aimed at recognizing mentions of named entities in web documents in Slavic languages, their normalization / lemmatization, and cross-language matching (Marcińczuk et al., 2017). The system also supported the annotation of the corpora constructed specially for specific tasks from the field of natural language processing e.g. Polish Coreference Corpus for definite descriptions annotation and Polish Spatial Texts corpus for the annotation of dynamic spatial expressions. It involved creation of dedicated annotation layers but, what is important, in these tasks the new module of the system (Annotation Agreement and ”2+1” annotation) was used for the first time, which significantly improved the time of preparation of annotated training and testing corpora.

5 Summary

Inforex system, as a part of CLARIN-PL infrastructure, is gradually developed. Although its initial role was to construct qualitative linguistic resources for various tasks from the field of natural language processing, recently it is also used by scientists for other purposes. We received an important and constructive feedback from users during and after workshops related to CLARIN-PL tools and resources. As users have different needs, we identified the common functionalities and implement them as soon as possible in order to boost their research tasks and provide new possibilities. We also challenged with the fact that many researches from the field of digital humanities are not experienced users of such systems and we made Inforex as easy and intuitive as possible.

Acknowledgments

Work financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education.

References

Dominik Bas, Bartosz Broda, and Maciej Piasecki. 2008. Towards Word Sense Disambiguation of Polish. In *Proceedings of the International Multiconference on Computer Science and Information Technology, {IMCSIT} 2008, Wisla, Poland, 20-22 October 2008*. IEEE, pages 73–78. https://doi.org/10.1109/IMCSIT.2008.4747220.

Kalina Bontcheva, Hamish Cunningham, Ian Roberts, Angus Roberts, Valentin Tablan, Niraj Aswani, and Genevieve Gorrell. 2013. Gate teamware: a web-based, collaborative text annotation framework. *Language Resources and Evaluation* 47(4):1007–1029.

Bartosz Broda, Michał Marcińczuk, Marek Maziarz, Adam Radziszewski, and Adam Wardyński. 2012. KPWr: Towards a Free Corpus of Polish. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of LREC'12*. ELRA, Istanbul, Turkey.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan, Niraj Aswani, Ian Roberts, Genevieve Gorrell, Adam Funk, Angus Roberts, Danica Damljanovic, Thomas Heitz, Mark A. Greenwood, Horacio Saggion, Johann Petrak, Yaoyong Li, and Wim Peters. 2011. *Text Processing with GATE (Version 6)*. http://tinyurl.com/gatebook.

Richard Eckart de Castilho, Eva Mujdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chris Biemann. 2016. A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH) at COLING 2016*. pages 76–84.

George Hripcsak and Adam S. Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *J. of Am. Medical Informatics Association* 12(3):296–298.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The sketch engine: ten years on. *Lexicography* .

Jan Kocoń, Michał Marcińczuk, Marcin Oleksy, Tomasz Bernaś, and Michał Wolski. 2015. Temporal expressions in polish corpus kpwr. *Cognitive Studies— Études cognitives* (15):293–317.

LCD. 2005. ACE (Automatic Content Extraction) English Annotation Guidelines for Events. Technical report, Linguistic Data Consortium.

M. Marcińczuk and M. Ptak. 2012. *Preliminary study on automatic induction of rules for recognition of semantic relations between proper names in Polish texts*, volume 7499 LNAI.

Michał Marcińczuk, Jan Kocoń, and Marcin Oleksy. 2017. Liner2 — a generic framework for named entity recognition. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*. Association for Computational Linguistics, Valencia, Spain, pages 86–91. http://www.aclweb.org/anthology/W17-1413.

Michał Marcińczuk, Marcin Oleksy, Tomasz Bernaś, Jan Kocoń, and Michał Wolski. 2015. Towards an event annotated corpus of polish. *Cognitive Studies— Études cognitives* (15):253–267.

Michał Marcińczuk, Michał Stanek, Maciej Piasecki, and Adam Musiał. 2011. Rich Set of Features for Proper Name Recognition in Polish Texts. In *SIIS 2011*. Springer.

Michał Mirosław Marcińczuk, Marcin Oleksy, and Jan Wieczorek. 2016. Towards recognition of spatial relations between entities for polish. *Cognitive Studies— Études cognitives* (16):119–132.

Małgorzata Marciniak, Agnieszka Mykowiecka, and Katarzyna Głowińska. 2010. Anotowany korpus dialogów telefonicznych. In Małgorzata Marciniak, editor, *Anotowany korpus dialogów telefonicznych*, Akademicka Oficyna Wydawnicza EXIT, Warsaw, chapter Anotacja korpusu LUNA–WOZ.PL, pages 217–230.

Michał Marcińczuk, Jan Kocoń, and Maciej Janicki. 2013. Liner2 – a customizable framework for proper names recognition for Polish. In Robert Bembenik, Lukasz Skonieczny, Henryk Rybinski, Marzena Kryszkiewicz, and Marek Niezgodka, editors, *Intelligent Tools for Building a Scientific Information Platform*, pages 231–253.

Michał Marcińczuk, Monika Zaśko-Zielińska, and Maciej Piasecki. 2011. Structure annotation in the polish corpus of suicide notes. In Ivan Habernal and Václav Matoušek, editors, *Text, Speech and Dialogue*, Springer Berlin Heidelberg, volume 6836 of *Lecture Notes in Computer Science*, pages 419–426.

Maciej Ogrodniczuk, Katarzyna Głowińska, Mateusz Kopeć, Agata Savary, and Magdalena Zawisławska. 2015. *Coreference in Polish: Annotation, Resolution and Evaluation*. Walter De Gruyter. http://www.degruyter.com/view/product/428667.

Adam Radziszewski and Maciej Piasecki. 2010. A Preliminary Noun Phrase Chunker for Polish. *Proceedings of the Intelligent Information Systems* pages 169–180.

Pavel Rychlý. 2007. Manatee/bonito - a modular corpus manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*. Masarykova univerzita, Brno, pages 65–70.

M. Zaśko-Zielińska. 2013. *Listy pożegnalne: w poszukiwaniu lingwistycznych wyznaczników autentyczności tekstu*. Quaestio. https://books.google.pl/books?id=QG60ngEACAAJ.

Figure 5: Summary of annotation agreement for a set of document

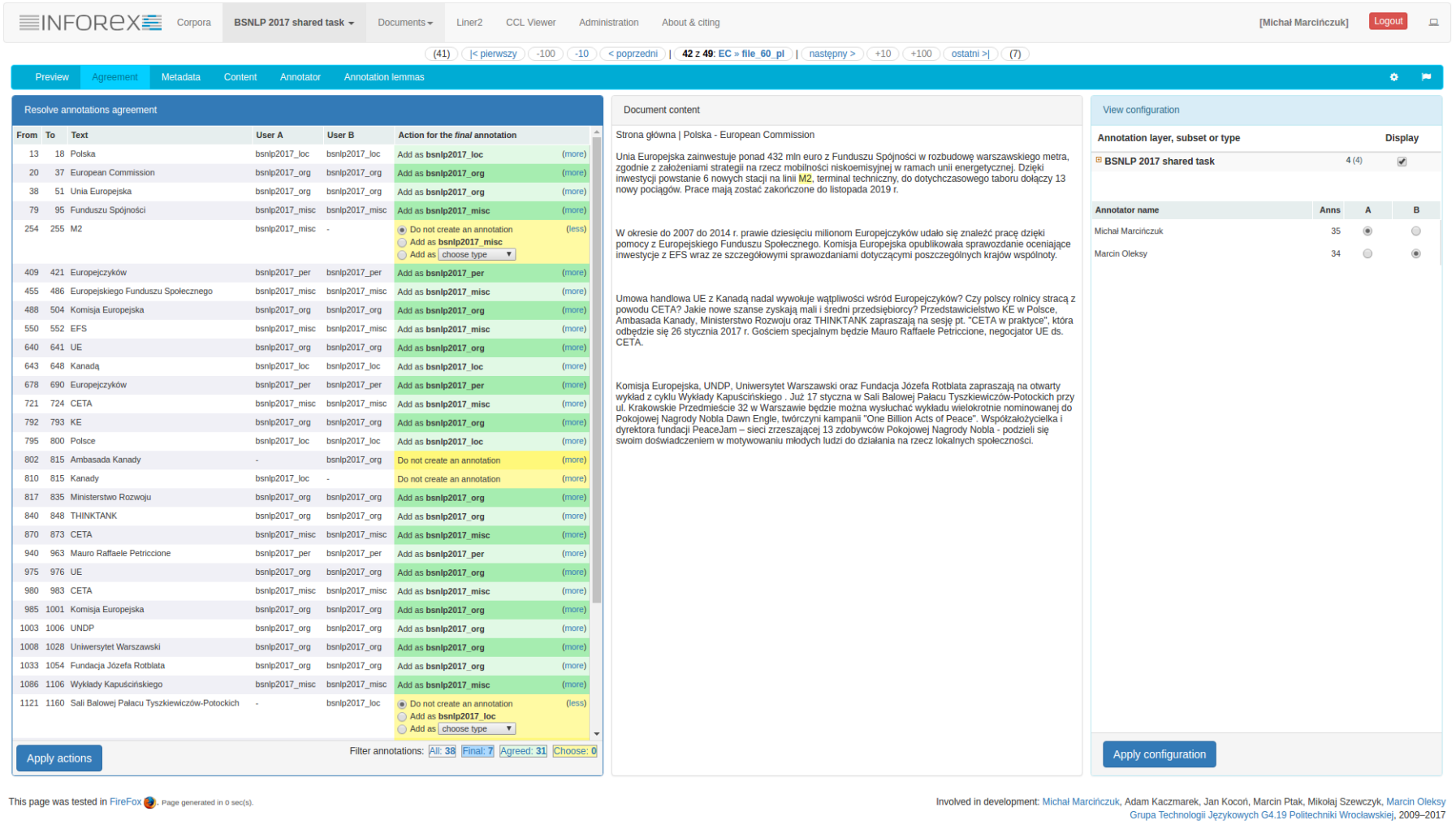


Figure 6: User agreement verification for a single document