SEJFEK — a Lexicon and a Shallow Grammar of Polish Economic Multi-Word Units

Agata Savary[1] *Bartosz Zaborowski*[2]
Aleksandra Krawczyk-Wieczorek[2] *Filip Makowiecki*[3]
(1) Université François Rabelais Tours, Laboratoire d'informatique, Blois, France
(2) Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland
(3) University of Warsaw, Poland
agata.savary@univ-tours.fr, bartosz.zaborowski@ipipan.waw.pl,
aleksandra.wieczorek@ipipan.waw.pl, f.makowiecki@gmail.com

Abstract
We present a large-coverage lexical and grammatical resource of Polish economic terminology. It consists of two alternative modules. One is a grammatical lexicon of about 11,000 terminological multi-word units, where inflectional and syntactic variation, as well as nesting of terms, are described via graph-based rules. The other one is a fully lexicalized shallow grammar, obtained by an automatic conversion of the lexicon, and partly manually validated. Both resources have a good coverage, evaluated on a manually annotated corpus, and are freely available under the Creative Commons BY-SA license.

Keywords: electronic lexicon, shallow grammar, Polish, economic terminology, language resources and tools.

1 Introduction

Terminology is one of important application domains of Natural Language Processing (NLP). Information extraction, text classification, automatic summarization, machine translation and other NLP fields can greatly support the exploitation of specialized texts by both experts and a large public. These processes heavily rely on identification and understanding of technical terms which are semantically rich linguistic units.

The basic facts about terms are that: (i) terminology is very productive: new terms are constantly created with the rapid advances of science and technology, (ii) most of them are nominal multi-word units (MWUs), (iii) many multi-word terms contain other, previously forged, terminological MWUs, e.g. *read-only memory* (*ROM*), *programmable ROM*, *erasable programmable ROM*, etc. The long tradition of terminological extraction shows that particularly interesting results can be obtained with hybrid approaches which combine statistical lexical association measures and shallow parsing (Smadja, 1993; Daille, 1996). Prevalent inflectional, syntactic and semantic variability of terminological MWUs calls for fine-grained representation of their linguistic properties (Jacquemin, 2001). Moreover the necessity of "looking inside" terminological MWUs, in order to recognize their nested structures, has been more recently recognized (Alex et al., 2007; Finkel and Manning, 2009).

While some work has been done in automatic processing of terminology for Slavic languages (Koeva, 2007; Mykowiecka et al., 2009), which are morphologically complex, relatively

few large-coverage NLP resources exist for automatic processing of terminology in these languages. Our work contributes to bridging this gap. We present *SEJFEK*, an NLP-oriented resource for Polish in the domain of economy. It consists of two alternative modules. One is a grammatical lexicon of about 11,000 terminological MWUs, where inflectional and syntactic variation, as well as nesting of terms, are described via fine-grained rules (cf. Sec. 2). The other one is a fully lexicalized shallow grammar, obtained by an automatic conversion of the lexicon, and manually validated (cf. Sec. 3).

2 Grammatical Lexicon of Polish Economic Phraseology

SEJFEK (*Słownik Elektroniczny Jednostek Frazeologicznych z EKonomii*)[1] was created as a grammatical lexicon of Polish economic phraseology. In this section we describe the scope of this resource, the data selection process, the formalisms and tools used for the lexicographic work, and the current contents of the lexicon.

2.1 Knowledge Sources

Constructing any lexical resource has to start with defining its precise scope. We have carried out some initial studies concerning the question which areas should precisely be considered as belonging to the domain of economy. Micro- and macroeconomy, banking, finance, economic policy, trade and international economics seemed undoubtedly relevant, while marketing, management and employment policy might be seen as borderline with respect to economy. We finally relied on the Resolution of the Central Commission for Degrees and Titles of June 23, 2006[2]. We have selected all domains, except commodity, considered in this official document as parts of economic sciences: economy, finance and management with their subdomains. Linguistically speaking, the terms to be included in the lexicon were to be multi-word nominal units with a reasonably fixed terminological meaning. Both common and proper nouns were considered relevant. Quantitatively speaking, the funding project allowed for the description of about 10,000 entries.

The collection of input material has been done mainly manually. The main lexicographer was an expert in linguistics with a thorough knowledge of economy, which greatly facilitated and enhanced the reliability of both the data selection and its grammatical description. Initially, input data were searched for in the following the Web sources:

- *Encyklopedia Zarządzania* 'Encyclopedia of Management' (`http://mfiles.pl`) constructed within a collaborative Wiki framework and containing (at the beginning of our project) about 4,000 terms. Many of them were simple words and had to be eliminated. Numerous relevant data were manually selected from tables and schemas.
- *Money.pl* (`www.money.pl/slownik`), *Bankier.pl* (`www.bankier.pl/slownik`) and *NBPortal.pl* (`http://www.nbportal.pl/pl/np/slownik`) – targeted but relatively small web lexicons.
- Official portals of Polish finance and political institutions, notably *Narodowy Bank Polski* 'Polish National Bank' (`www.nbp.pl`), *Ministerstwo Finansów* 'Ministry of Finance' (`www.mf.gov.pl`), and *Giełda Papierów Wartościowych w Warszawie* 'Warsaw Stock Exchange' (`www.gpw.pl`). Manual browsing of articles and guides allowed to extract additional terms, as well as some proper names, e.g. the list of companies

[1] `http://zil.ipipan.waw.pl/SEJFEK`

[2] Uchwała Centralnej Komisji do spraw Stopni i Tytułów z 23.06.2003

listed in the Warsaw Stock Exchange, financial and political institutions, economic programs, and the *Polska Klasyfikacja Działalności* 'Polish Classification of Activities'.
- Economic and financial services of major Polish web portals (`onet.pl, wp.pl, gazeta.pl, forsal.pl`). Their texts showed a rather low density of economic terms as they were mainly addressed to non specialists.

An attempt was made to extract candidate terms automatically from corpora with a Polish Web crawler and collocation finder *Kolokacje*[3], which however yielded few valuable results. In view of this experiment we think that automatic terminological extraction might greatly benefit from high quality lexical and grammatical resources, such as those described below.

The list of terms selected from the web was further completed with data from indexes of traditional printed economic lexicons and manuals. Those were chosen from bibliographical lists recommended for students of economy and management at the University of Warsaw and included: (Samuelson and Nordhaus, 2003), (Samuelson and Nordhaus, 1998), (Głuchowski and Szambelańczyk, 1999), (Wernik, 2007), (Michoń, 1991), (Rynarzewski and Zielińska-Głębocka, 2006), (Treder, 2005), (Kuciński, 2009), (Chow, 1995), (Śnieżek, 2004), (Michalski, 2003), (Black, 2008), and (Smullen and Hand, 2008). Some terminology dedicated to European integration was found in (Rzewuska et al., 2001).

2.2 Formalism and Tool

After selecting the economic MWUs to be included in the lexicon, their grammatical description was done within *Toposław* (Marciniak et al., 2011), the lexicographic framework initially meant for the development of lexical resources of Polish proper names (Savary et al., 2009). This platform offers a user-friendly graphical interface encompassing three core components: (i) *Morfeusz*, the morphological analyzer and generator of Polish simple words, (ii) *Multiflex* (Savary, 2009), a graph-based generator of inflected forms of multi-word units, (iii) a graph editor stemming from *Unitex* [4], a multilingual corpus processor.

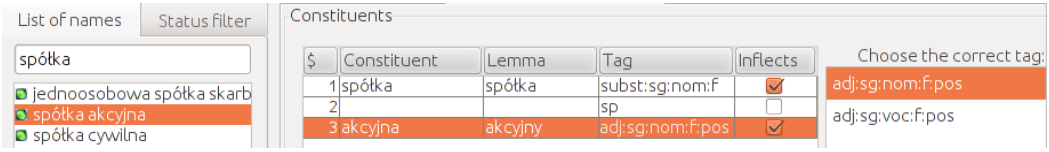

Figure 1: Describing the components of *spółka akcyjna* 'joint-stock company' in Toposław.

The grammatical description of MWUs in Toposław is organized in two steps. Firstly, the internal structure of each term is modeled in that the MWU is divided into numbered tokens, each token is analyzed by Morfeusz and disambiguated manually by the lexicographer. The components which can vary during the inflection of the whole MWU are also marked. Fig. 1 shows the internal structure of *spółka akcyjna* 'joint-stock company'. Three components are delimited: (i) *spółka* 'company' – a substantive (*subst*) in singular (*sg*), nominative (*nom*), feminine (*f*), (ii) a blank space, (iii) *akcyjna* 'joint-stock' – an adjective (*adj*) in singular, ambiguous between nominative and vocative (*voc*), feminine, positive degree (*pos*). The first and the third component can inflect when the whole MWU is inflected.

[3] `http://www.mimuw.edu.pl/polszczyzna/kolokacje/index.htm`
[4] `http://www-igm.univ-mlv.fr/~unitex/`

Secondly, the MWU as a whole is assigned the proper *inflection graph* which describes the generation of its inflected forms and variants. Fig. 2 shows the inflection graph for the MWU analyzed in Fig. 1. The leftmost triangle represents the entry point of the graph, while the encircled square shows its exit. The numbered boxes correspond to constituents of the name (words, spaces, punctuation or sub-compounds). The arrow-laden lines that connect the boxes represent various paths which can be used while generating the inflected forms of a name. Here, the bottom-most path describes the acronymic variant *S.A.* The formulae inside boxes consist of constituents' indexes and equations on morphological constants and variables. These equations impose constraints on the inflection, variation and agreement of constituents. For example, the equations containing constants such as $Init = dot$ and $LetterCase = first\_upper$ mean that only the capitalized initial letter of the current component is taken, followed by a dot. The equations containing variables, $Case = \$c$ and $Number = \$n$, allow the component to inflect for case and number. When these variables reoccur on the same path the respective components must agree, as in the case of component \$3 in the upper path of Fig. 2. The formulae appearing below paths determine the features of the inflected forms of the whole compound as a function of the features of its constituents. Here, the form resulting from each path inherits its gender from the first constituent and has the conforming case and number ($Case = \$c; Gen = \$1.Gen; Nb = \$n$).

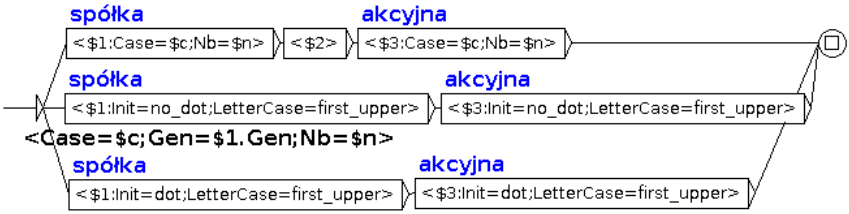

Figure 2: Inflection graph for *spółka akcyjna* 'joint-stock company' in Toposław.

When applying the graph in Fig. 2 to the MWU in Fig. 1 we obtain the set of all inflected forms shown in Tab. 1.

Inflected forms	Morphological features	Inflected forms	Morphological features
spółka akcyjna SA S.A.	subst:**sg:nom**:f	spółki akcyjne SA S.A.	subst:**pl:nom**:f
spółki akcyjnej SA S.A.	subst:**sg:gen**:f	spółek akcyjnych SA S.A.	subst:**pl:gen**:f
spółce akcyjnej SA S.A.	subst:**sg:dat**:f	spółkom akcyjnym SA S.A.	subst:**pl:dat**:f
spółkę akcyjną SA S.A.	subst:**sg:acc**:f	spółki akcyjne SA S.A.	subst:**pl:acc**:f
spółką akcyjną SA S.A.	subst:**sg:inst**:f	spółkami akcyjnymi SA S.A.	subst:**pl:inst**:f
spółce akcyjnej SA S.A.	subst:**sg:loc**:f	spółkach akcyjnych SA S.A.	subst:**pl:loc**:f
spółko akcyjna SA S.A.	subst:**sg:voc**:f	spółki akcyjne SA S.A.	subst:**pl:voc**:f

Table 1: Inflected forms of *spółka akcyjna* 'joint-stock company'.

The Multiflex graph formalism allows also to represent embedding of MWUs within other MWUs. Fig. 3 shows the components of a name of a bank, *Bank BPH Spółka Akcyjna*, with the nested MWU discussed above. Note that *Spółka Akcyjna* is analyzed here as a unique multi-word component with number 5. Toposław supports the manual description of embedding by automatically matching the nesting and the nested entries.

Nested structures allow to establish links between different entries of the lexicon, which can be later exploited in semantic processing of texts. Moreover, the inflection graphs are simpler if nesting is taken into account and their number is lower. Fig. 4 shows the graph

for the entry in Fig. 3. The upper path corresponds to all inflected forms of the entry (in singular only), with components $1 and $5 agreeing in case, and with the last component taking any of its possible variants (*Spółka Akcyjna*, *S.A.* or *SA*). The lower path describes the elliptical variant *Bank BPH* and its inflection for case. If the sub-term *Spółka Akcyjna* was not delimited as nested then the corresponding graph would have to be much more complex. It would have to explicitly contain all three paths of the graph from Fig. 2.

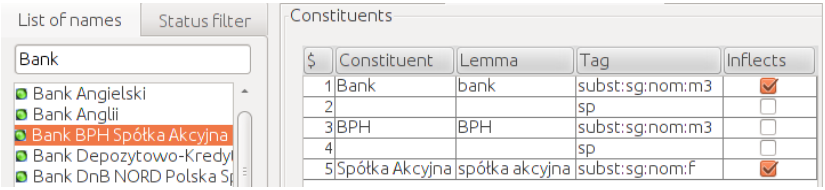

Figure 3: Describing a nested multi-word component in *Bank BPH Spółka Akcyjna* 'BPH Joint-Stock Bank'.

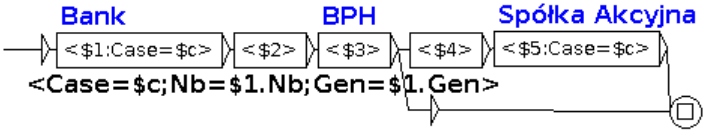

Figure 4: Inflection graph for *Bank BPH Spółka Akcyjna* 'BPH Joint-Stock Bank' with a nested component.

The result of the application of the graph in Fig. 4 to the entry in Fig. 3 is shown in Tab. 2. Note that the nested MWU *Spółka Akcyjna* is a graphical variation (with uppercase initials) of its lemma *spółka akcyjna*. The variation of this kind is automatically reproduced by Multiflex during the inflection process.

Inflected forms				Morphological features
Bank BPH Spółka Akcyjna	Bank BPH SA	Bank BPH S.A.	Bank BPH	subst:sg:**nom**:m3
Banku BPH Spółki Akcyjnej	Banku BPH SA	Banku BPH S.A.	Banku BPH	subst:sg:**gen**:m3
Bankowi BPH Spółce Akcyjnej	Bankowi BPH SA	Bankowi BPH S.A.	Bankowi BPH	subst:sg:**dat**:m3
Bank BPH Spółkę Akcyjną	Bank BPH SA	Bank BPH S.A.	Bank BPH	subst:sg:**acc**:m3
Bankiem BPH Spółką Akcyjną	Bankiem BPH SA	Bankiem BPH S.A.	Bankiem BPH	subst:sg:**inst**:m3
Banku BPH Spółce Akcyjnej	Banku BPH SA	Banku BPH S.A.	Banku BPH	subst:sg:**loc**:m3
Banku BPH Spółko Akcyjna	Banku BPH SA	Banku BPH S.A.	Banku BPH	subst:sg:**voc**:m3

Table 2: Inflected forms of *Bank BPH Spółka Akcyjna* 'BPH Joint-Stock Bank'.

A lexicon in Toposław can be exported to a Multiflex-compatible textual format as shown in Ex. (1)–(2). The final information (inside parentheses) is the inflectional graph's name. Toposław partly constraints this name so as to fit the syntactic structure of the assigned entries. E.g., `NC-O_O` means that the structure is a nominal compound with two inflected (***O**dmienny* in Polish) components, while `NC-O_N_O` suggests two inflected (here: *Bank* and *Spółka Akcyjna*) and one non-inflected (***N**ieodmienny* in Polish, here: *BPH*) component. The remaining part of the graph name is freely chosen by the lexicographer, who may fix his own convention. Here, `nb-inv` suggests that the entry is invariable in number.

(1) spółka(spółka:subst:sg:nom:f) akcyjna(akcyjny:adj:sg:nom:f:pos),subst(NC-O_O-SA)

(2) Bank(bank:subst:sg:nom:m3) BPH(BPH:subst:sg:nom:m3)
{Spółka Akcyjna}(spółka akcyjna:subst:sg:nom:f),subst(NC-O_N_O-nb-inv-SA)

2.3 Contents of the Lexicon

Tab. 3 shows the current state of SEJFEK. Complete entries are those whose inflected components are known to Morfeusz, thus the generation of the inflected forms for these entries could be fully performed. Conversely, problematic entries are those containing unknown components, mostly proper names and inflected acronyms (cf. the first dot in Sec. 2.4).

MWU lemmas		**Inflected forms**	**Graphs**
Complete	Problematic		
11,211	141	146,861	293

Table 3: Contents of the lexicon.

The high number of inflection graphs results from a big variety of syntactic structures typical for technical terms, as well as from their high degree of variability (acronyms, ellipses, word order change, restrictions in number inflection, etc.). Tab. 4 shows statistics of graph assignment. The first 6 lines concern the most frequently assigned graphs, as well as examples of different internal structures of the assigned entries. The agreement structures of type *SubstAdj* and *AdjSubst* as well as the government structures of type $SubstSubst_{gen}$ are the most frequent ones in both nesting and nested terms. For instance *[[czytnik elektroniczny] [kodów kreskowych]]* 'barcode reader (lit. [[electronic reader] of [bar codes]])' has the internal structure of type $Subst(SubstAdj)Subst_{gen}(Subst_{gen}Adj_{gen})$.

Note that embedding of terms is considered on a semantic rather than syntactic basis. For instance the term *teoria powiązań pionowych i poziomych między firmami* 'theory of vertical and horizontal links between firms' can be syntactically parsed into a constituency tree of depth 6. However it has a flat semantic structure in SEJFEK since none of its substrings is an economic term on its own.

2.4 Interesting Problems

We give several examples of problems that had to be faced by the lexicographer during morphosyntactic description of terms in SEJFEK:

- **Unknown words** As shown in Section 2.2 the inflection of a MWU consists essentially in combining the proper inflected forms or variants of its components. Consequently, both the morphological analysis and generation is required for the components which vary during the inflection of the whole MWU. Some components were unknown to Morfeusz at the period of the SEJFEK development, notably foreign proper names (*David Hume*, *Davida Hume'a*), foreign common words which inflect in Polish (*Allianz Polska*, *Allianzu Polska*), inflected acronyms (*FAM S.A.*, *FAM-u S.A.*), Polish technical terms (*doktryna libertarianistyczna* 'libertarianist doctrine') and Polish derivation forms (*konkurencja pozacenowa* 'non-price rivalry', *popyt zagregowany* 'aggregate demand'). In order to obtain the correct inflection of the latter cases, problematic derivatives were frequently divided into several known tokens (*poza+cenowa*). Sometimes this division was artificial (*z+agregowany*) and should be eliminated as soon as Morfeusz' dictionary gets sufficiently enlarged.
- **Grammatical homonyms** Some components known to Morfeusz were subject to shift in gender while appearing in economic terms. For instance, the first component in *estymator odporny* 'robust estimator' was analyzed as human masculine noun ($m1$

gender) but it has the human inanimate ($m3$) gender.

- **Unclear inflection paradigm** The lexicographer frequently faced a lack of evidence with respect to the inflection of some proper names, particularly those containing foreign components. For instance *Allianz Polska* might remain unaltered in genitive or might have its first component inflected: *Allianzu Polska*.
- **Productive structures** Some institution names followed a very productive schema, e.g. *Urząd Skarbowy w Białymstoku*, *Urząd Skarbowy w Bydgoszczy*, etc. 'Treasury Office in Białystok/Bydgoszcz/...'. These names were not systematically listed in the lexicon as they would much more conveniently be expressed by regular expressions.

Graphs	**Uppermost syntactic structure**		**Examples**	**Assigned entries**
	Agreement	**Government**		
NC-O_O	S Adj Adj S		*spółka akcyjna* *zloty spadochron, agresywna [zmiana cen]*	2,573
NC-O_N-nb-inv		S S_{gen}	*krzywa Beveridge'a, [ryzyko inwestycyjne] obligacji, demonetyzacja [zagranicznych [środków płatniczych]]*	1,482
NC-O_N		S S_{gen}	*centrum rozliczeń, czynnik [kreacji podaży], [kryterium operacyjne] denominacji, analiza [polityki [wydatków publicznych]], [[czytnik elektroniczny][kodów kreskowych]], podstawa [wymiaru [składek [ubezpieczeń społecznych]]]*	1,320
NC-O_O-nb-inv	S Adj Adj S		*aktywa niematerialne, [produkt narodowy brutto] realny wtórne [ryzyko płynności]*	1,156
NC-O_N_N-nb-inv		S S_{gen} S_{gen} S Prep S_{gov}	*częstotliwość dokonywania zakupu* *egzekucja z [wynagrodzenia za pracę]* *[poziom dobrobytu] w [skali krajowej]*	662
NC-O_O-ord	S Adj Adj S		*dotacja bezpośrednia, [dług ekonomiczny] użytkowy* *lokalne [dobro publiczne]*	551
Others			*teoria powiązań pionowych i poziomych między firmami*	3,064
Total				11,352

Table 4: Distribution of graphs and variability of internal structures in assigned entries. The following codes are used: nominal compound (NC), variable component (O), invariable component (N), invariability in number ($nb{-}inv$), variability in order (ord), substantive (S), substantive in genitive (S_{gen}), substantive in a case governed by the preposition (S_{gov}), and adjective (Adj).

3 From Lexicon to Shallow Grammar

A grammatical lexicon such as SEJFEK is currently mainly generation-oriented, i.e. the semantics of inflection graphs was designed in view of automatic generation of all inflected forms and variants of a MWU. The resulting list of over 146,000 forms may be used in particular for matching terms in the process of a MWU-aware morphological analysis of a text, as is the case e.g. in the *Unitex* corpus processor (Paumier, 2008). However this approach, although simple and straightforward, has the disadvantage of not being able to transmit the data about the internal, syntactic or semantic, structure of a recognized MWU to further stages of linguistic processing. Therefore, we wished to experiment with the feasibility of transforming this rich lexical resource into a shallow grammar. The grammatical formalism chosen for this experiment is *Spejd* (Przepiórkowski, 2008; Przepiórkowski and Buczyński, 2007; Zaborowski, 2012).

3.1 Spejd Formalism

Spejd's input is a morphologically analyzed text, in which each token possibly gets several morphosyntactic interpretations. While tagging and (partial) parsing are usually done as separate processes, Spejd combines them into one parallel process: it allows to simultaneously disambiguate and build syntactic structures within a single rule. A *Spejd grammar* is a cascade of regular grammars (each of the rules is a separate grammar). A rule falls into 2 parts – a matching part and a list of operations — the former is divided into sections.

The *matching part* specifies a pattern of tokens and/or syntactic structures, as well as their (optional) context. The *Match* section is a regular expression over token specifications. In our case each rule will represent one MWU term, thus regular expressions come down to sequences. A specification of a token consists of constraints on its morphosyntactic features. A constraint contains an attribute name, a comparison operator and a regular expression specifying the desired value. Multiple requirements for a single token are connected with conjunction (**&&**) which applies at the level of a single interpretation. In our case the most useful comparison operators are ~ and ~~. The former means that there is at least one interpretation of the token which satisfies the constraint. The latter ensures that all its interpretations do alike. For a non ambiguous token both operators are equivalent.

Ex.(3) shows a sample rule whose matching part matches two tokens. The first one is a noun (`pos~"subst"`) and has the lemma *spółka* (`base~"spółka"`, `/i` stands for case-insensitive). The second one must be an adjective and must have the (case-insensitive) lemma *akcyjny* . The capital letters `A` and `B` enable referring to particular tokens from the second part of the rule. The additional sections of the matching part (e.g. a context specification), which are not used here, can be built in a similar way.

(3)
```
Rule  "syntok Spółka Akcyjna"
Match:  A[base~"spółka"/i && pos~subst] B[base~"akcyjny"/i && pos~adj];
Eval:  unify(case gender number, A,B);
       leave(base~~"spółka", A); leave(pos~~"subst", A);
       leave(base~~"akcyjna", B); leave(pos~~"adj", B);
       word(A, , "Spółka Akcyjna");
```

The second part of a rule consist of a *list of operations* preceded by the keyword `Eval`, and executed in the order they appear in the list. Some of them, e.g. *unify*, evaluate to a Boolean value (similarly to predicates in PROLOG). When an operation evaluates to *false*, the execution is broken (like in PROLOG) but the changes made by previous operations are not rolled back (contrary to PROLOG).

In Ex.(3) the *unify* operation checks for agreement in case, gender and number between tokens `A` and `B`. If these tokens have no interpretations with the same values on those attributes, the operation returns false and the execution of the list breaks. Otherwise all combinations of interpretations which violate agreement are removed and the evaluation continues. The *leave* operations remove all those interpretations of tokens `A` and `B` which have lemmas different from *spółka* and *akcyjna* or parts of speech different from `subst` and `adj`, respectively. The last operation (*word*) builds a syntactic word consisting of all matched tokens with morphosyntactic interpretations copied from the token `A` and lemma *"Spółka Akcyjna"*. As a result, the rule matches all 14 inflected forms shown at the first position of each line in Table 1, as well as their capitalized variants.

3.2 Conversion Methodology

In the original form, the lexicon is represented by a list of entries annotated by a set of graphs. Since the semantics of graphs is complex and not easily transformable into a grammar, we base our conversion on a textual representation of the lexicon, as in Ex. (1)–(2). It discards the detailed information contained in graphs but simplifies further automatic processing and still allows to perform analysis. In some rare cases this approach led us to problems described in Section 3.3.

3.2.1 The conversion algorithm

The main assumptions for the conversion algorithm are the following:

- For each term appearing in the lexicon, the grammar should build a syntactic word.
- The word's morphosyntactic features are derived from its headword.
- The correct recognition of terms should be ensured by unification of inflection features.
- Nested terms should be properly represented as nested syntactic words.

The conversion relies on the term's general structure (shown in the name of its inflection graph, cf. Sec. 2.2). Ex. (4) shows the Spejd rule resulting from converting the lexicon term with structure `O_N_O` from Ex. (2). The matching pattern is created with constraints on the word's: (i) lemma (case-insensitive), POS, and negation value (for participles only) if the component is **inflected** (here: *Bank* and *Spółka Akcyjna*; the latter is a nested term recognized previously by a dedicated rule), (ii) orthographic form (case-insensitive) if it is **uninflected** (here: *BPH*). We have also experimented with allowing a formally uninflected word to be plural in order to cover cases such as *jakość produktu/produktów* 'quality of product(s)'. This property may over-generate, but proves useful for the purpose of analysis.

(4)
```
Rule  "syntok Bank BPH Spółka Akcyjna"
Match:  A[base~"bank"/i && pos~subst] [orth~"BPH"/i]
          B[base~"Spółka Akcyjna"/i && pos~subst];
Eval: unify(case, A,B);
      leave(base~~"bank"/i, A); leave(pos~~"subst", A);
      leave(base~~"spółka akcyjna"/i, B); leave(pos~~"subst", B);
      word(A, , "Bank BPH Spółka Akcyjna");
```

As explained in Sec. 3.1, the *Eval* section of a rule should: (i) ensure the term is correctly recognized, (ii) disambiguate it morphosyntactically, (iii) build a syntactic word. Task (i) is performed for most terms by a naive approach: unification in case, number and gender is required between all inflected components, as in Ex. (3). For some rare exceptions, as in Ex. (4), the unification is limited to the case (cf. Sec. 3.3). Task (ii) is performed by *leave* clauses which conserve for each inflected component only those interpretations whose lemmas and POSs match the morphosyntactic annotation in the lexicon (here: *bank* and *subst* for *Bank*, and *spółka akcyjna* and *subst* for *Spółka Akcyjna*). Task (iii) is done by the 3-argument "copying" version of the *word* action: the morphosyntactic features for the resultant syntactic word are copied from the headword (here: *Bank*) while the resulting lemma is constructed by simple concatenation of component forms (here: *Bank BPH Spółka Akcyjna*). The headword is determined according to the following rules:

- inflected elements take precedence over non-inflected ones,

- nouns (*subst* and *ger*) have a higher priority than adjectives (*adj*, *pact* and *ppas*),
- the case of the headword in the MWU's lemma must be nominal,
- if the above rules select more than one element, the left-most one is selected.

3.3 Problems with Conversion

As mentioned above, only the textual export form of the lexicon was used for conversion, which was sufficient in the majority of cases but provoked three main problems. Firstly, and most importantly, the morphosyntactic variants not expressed on the level of a graph's name could not be taken into account. In particular, word order change, elliptical variants and acronyms, as those described by the graph in Fig. 2, are currently not recognized.

Secondly, the general rule of imposing number, case and gender agreement of all inflected components (cf. Sec. 3.2) failed in appositions and coordinations, where several components may agree in case but usually only one of them is the headword. In Ex. (2) *Bank* is in masculine inanimate ($m3$), and *Spółka Akcyjna* in feminine (f) but both agree in case. In Ex. (5)[5] the first and the third constituent differ both in gender and in number but they still agree in case. Such cases were manually marked in the lexicon before conversion and the corresponding Spejd rules were tuned so as to perform case unification only, as shown in Ex. (4). We think that an automated procedure might help detect such apposition and coordination cases and restrict agreement to case accordingly. Special care must however be taken if nouns are accompanied by adjectival modifiers. Moreover some appositions may even exclude case agreement of nouns, as in *Allianz Polska*, *Allianzu Polska*, etc.

(5) kapitał(kapitał:subst:**sg**:nom:**m3**) i rezerwy(rezerwa:subst:**pl**:nom:**f**) 'capital and reserves'

(6) *old entry*: funkcja **Cobba-Douglasa(:qub)**,subst(NC-**O__N**-nb-inv)
new entry: funkcja **Cobba(:qub)-(:interp)Douglasa(:qub)**,subst(NC-**O__NNN**-nb-inv)
'Cobb-Douglas function'

(7) *old entry*: Runda **Kennedy'ego**(Kennedy:subst:sg:gen:m1) 'Kennedy Round'
new entry: Runda **{Kennedy'ego}**(Kennedy:subst:sg:gen:m1)
added rule:
```
Match:  [orth~"kennedy"/i] ns [orth~"'"/i] ns [orth~"ego"/i];
Eval:  word(subst:sg:gen:m1, "kennedy");
```

Thirdly, the tokenization conventions might differ between the lexicon and the grammar. In Morfeusz, in which tokenization is inherent in morphological analysis, some sequences with hyphens or apostrophes, such as *Cobba-Douglasa* or *Kennedy'ego*, were seen as unique tokens because they can be compound names or inflected forms of one-word names. Spejd always divides them into 3 tokens. Thus, entries such as in the first lines in Ex. (6)[6]–(7) could not yield an operational Spejd rule and had to be transformed as shown in the lines below. Additionally, an extra rule for the new nested term *Kennedy'ego* had to be created in Spejd, as shown at the bottom of Ex. 7.

3.4 Conversion as a validation

During the automatic lexicon-to-grammar conversion some errors and inconsistencies could be spotted and corrected in the grammar (their correction in the lexicon will be done

[5] For readability reasons only the relevant parts of the lexicon entries are shown in Examples (5)–(7).

[6] The `qub` label is a dummy POS chosen for the obviously nominal names *Cobb* and *Douglas* due to the fact that these names are currently unknown to Morfeusz. Since they never vary in this MWU they do not have to be fully analyzed for the sake of inflection of the MWU.

shortly). Below we give examples of the most frequent errors[7]:

- Failing markup of a nested term, despite the existence of a lexicon entry for the subterm, cf. Ex. (8). These errors concerned about 1,000 entries. If they were not corrected Spejd would completely fail to recognize these terms since it applies shorter rules first. The rule for a nested term such as *działalnością gospodarczą* would fire first, it would create a syntactic word, and its components would no longer be recognizable separately by the larger rule. Such errors were automatically corrected by a naive script which searched for common sequences of single word lemmas through all the terms in lexicon. Some remaining problems were corrected manually.
- Missing base entry for a nested term, cf. Ex. (9). This problem could be solved either by separating the components of the nested term or generating a new rule for it. The latter solution was applied. Since the detailed characteristics of the nested term were not easy to determine in a general case, a simplified ruled was created which only applied to the particular inflected form.
- Redundant plural entries, cf. Ex. (10). Other entries for the same terms, with a lemma in singular, already allowed inflection for number. The redundant entries were eliminated.
- Erroneous morphosyntactic features or lemma of a component due to grammatical syncretism, as in Ex. (11)–(12).
- Inconsistence if the graph name wrt. the entry's structure, cf. Ex. (13).
- Typographical mistakes, cf. Ex. (14).

(8) działalność(działalność:subst:sg:nom:f) gospodarcza(gospodarczy:adj:sg:nom:f:pos)
∗kierowanie **działalnością(działalność:...) gospodarczą(gospodarczy:...)**
kierowanie **{działalnością gospodarczą}(działalność gospodarcza:...)**
'business management'

(9) wyliczanie {agregatów monetarnych}(agregat monetarny:subst:pl:gen:m3)
'monetary aggregate estimation'
∗*missing entry*: agregat(agregat:...) monetarny(monetarny:...)
added rule:
```
Match:  [orth "agregatów"/i] [orth "monetarnych"/i];
Eval:  word(subst:pl:gen:m3, "agregatów monetarnych");
```

(10) **zasada**(zasada:subst:**sg**:nom:f) rachunkowości,subst(NC-O_N)
∗**zasady**(zasada:subst:**pl**:nom:f) rachunkowości,subst(NC-O_N-nb-inv)
'accountancy rules'

(11) ∗cechy(cecha:subst:**sg:gen**:f) demograficzno-społeczne pracowników
cechy(cecha:subst:**pl:nom**:f) demograficzno-społeczne pracowników
'demografically-social features of employees'

(12) ∗BIG Bank Gdański(**Gdańsk**:subst:pl:nom:m3)
BIG Bank Gdański(**gdański**:adj:sg:nom:m3) 'BIG Bank of Gdańsk'

(13) ∗krajowa {akcja kredytowa},subst(NC-O_**N**)
krajowa {akcja kredytowa},subst(NC-O_**O**) 'national credit action'

(14) ∗konkurencja poza(poza:qub)**ceno(cena:subst:sg:voc:f)**
konkurencja poza(poza:qub)**cenowa(cenowy:adj:sg:nom:f:pos)** 'non-price competition'

[7]For readability reasons only the relevant parts of the lexicon entries are shown in Examples (8)–(14). Each incorrect entry is preceded by an asterisk (∗).

3.5 Contents and Output of the Grammar

The Spejd grammar obtained by the SEJFEK lexicon conversion counts 11,266 rules. As many as 3,205 rules contain nested terms. Only 59 rules required human correction since they limit the unification of inflected components to case agreement only.

(15)
```
<syntok rule="syntok␣Bank␣BPH␣Spółka␣Akcyjna">
  <orth>Bankiem BPH Spółką Akcyjną</orth>
  <lex><base>Bank BPH Spółka Akcyjna</base><ctag>subst:sg:inst:m3</ctag></lex>
  <tok><orth>Bankiem</orth>
       <lex><base>bank</base><ctag>subst:sg:inst:m3</ctag></lex>
  </tok>
  <tok><orth>BPH</orth>
       <lex><base>BPH</base><ctag>subst:sg:nom:m3</ctag></lex>
       <lex><base>BPH</base><ctag>subst:sg:gen:m3</ctag></lex>...
  </tok>
  <syntok rule="syntok␣spółka␣akcyjna"><orth>Spółką Akcyjną</orth>
    <lex><base>spółka akcyjna</base><ctag>subst:sg:inst:f</ctag></lex>
    <tok><orth>Spółką</orth>
         <lex><base>spółka</base><ctag>subst:sg:inst:f</ctag></lex>
    </tok>
    <tok><orth>Akcyjną</orth>
      <lex><base>akcyjny</base><ctag>adj:sg:inst:f:pos</ctag></lex>
      <lex disamb="0"><base>akcyjny</base><ctag>adj:sg:acc:f:pos</ctag></lex>
    </tok></syntok></syntok>
```

Ex. (15) shows a simplified fragment of a Spejd output processed by the rule in Ex. (4). Each ⟨`syntok`⟩ element encodes a syntactic word. Nesting of syntactic words is determined by the ordering of grammar rules in the cascade, which is automatically deduced from lexicon entries. The ⟨`tok`⟩ elements reflect the input tokens. Morphosyntactic interpretations are encoded as ⟨`lex`⟩ elements. Note, that one of them (marked by the `disamb="0"` attribute) has been eliminated here by the *unify* action in Ex. (3) since it violates the case agreement.

4 Evaluation

In order to perform an evaluation of both the lexicon and the grammar we have prepared a manually annotated corpus of economic texts. It consists of fragments of the *plWikiEcono* corpus[8] containing Polish Wikipedia articles assigned to Wikipedia categories and subcategories in economy[9]. Because Wikipedia articles are of encyclopedic nature the density of technical terms they contain is very high (in comparison to economic newspapers and magazines or Wikinews). Thus, these texts seem particularly well suited for evaluating targeted lexical and grammatical resources like ours.

Wikipedia articles	Tokens	Compound terms		
		Occurrences		Unique forms
		Nouns	Adjectives	
191	220,905	11,106	11	6,805

Table 5: Statistics of the evaluation corpus consisting of Wikipedia economic articles

The corpus annotation has been performed by one annotator within the GATE platform (Wilcock, 2009). The annotation schema was rather simple: contiguous sequences of words

[8] `http://bach.ipipan.waw.pl/wiki/zil/Korpus%20plWikiEcono`
[9] `http://pl.wikipedia.org/wiki/Kategoria:Ekonomia`

judged as multi-word economic terms were to be tagged as such and their syntactic category was to be indicated. Only two categories proved relevant: *economic compound noun* and *economic compound adjective*. The annotator was neutral with respect to the project, i.e. she had been involved neither in creation of the lexicon, nor in its conversion to grammar. She had a deep linguistic knowledge but only a common knowledge of economy, which may partly bias the quality of the annotation. Tab. 5 resumes the contents of the resulting evaluation corpus.

In order to compare the lexicon approach and the grammar approach we automatically annotated the evaluation corpus by means of both methods. Both of them were applied within the Spejd framework but involving different modules. For the lexicon approach, we used the list of all inflected forms and variants of the lexicon terms. Spejd's dictionary module used this list for straightforward term matching in the corpus. The dictionary module built syntactic words so as to preserve the nesting structure of terms. The grammar approach involved the main (grammar) module of Spejd. It generated similar structures in the output — syntactic words with preserved nesting structure, as shown in Ex. (15) — but using the grammar for searching terms. It additionally performed a partial disambiguation, which was not done in the case of the lexical method.

The evaluation consisted in the comparison of the original annotation of the corpus and the automatically generated annotation produced by each method. Since we searched for multi-word terms, we used not only the standard binary measure (score 1 if the precise term was found, 0 otherwise), but also a weak correctness measure. The latter was based on accuracy of BIO-type (Begin-Inside-Outside) tags in the scope of each term and of its 1-word left and right context. The 11,117 terms present in the evaluation corpus yielded about 47,500 BIO tags extracted in this way (with an average of 4.27 tags per term).

Consider for instance the three-word manually tagged term in the sequence *niedawna [krajowa akcja kredytowa] była* 'recent [national credit action] was', whose corresponding tag sequence is `O-B-I-I-O`. If an automatic annotation yields `O-B-I-O-O`, it gets the score 4/5 (4 out of 5 BIO tags match). Similarly, for `B-I-I-O-O` the score is 1/5. For the exact match (`O-B-I-I-O`) this measure gives 1, which is equal to the standard binary measure.

method	**correctness**	**weak correctness**	**false positives**
lexicon	36.32%	64.66%	0.12%
lexicon (case insensitive)	41.43%	68.14%	0.21%
grammar	42.01%	68.45%	0.13%

Table 6: Evaluation results of the lexicon and the grammar.

The evaluation scores are presented in Tab. 6. Both approaches give very similar results. A notable difference appears only if the inflected lexicon is applied in a case-sensitive manner (the grammar is case-insensitive by default) since it results then in many false negatives e.g. at the beginning of a sentence or in article titles. This difference can be toned down by case-insensitive searching for lexicon terms at the cost of a slightly larger amount of false positives. In any case the percentage of false positives is extremely low. They result mostly from an uncertain terminological status of some MWUs (*państwo członkowskie* 'member state'), from some minor corpus annotation errors (non annotated *prawo poboru* 'rights issue') or from overlapping terms ($[_1$*wartość nominalna* $[_2$*banknotów*$]_1$ *w obiegu*$]_2$

'nominal value of currency banknotes'). This low number of false positives may be seen as an evidence of a high quality of the corpus annotation. Namely, almost each term which was included in the lexicon by the linguistics/economy expert and which appeared in the corpus was correctly spotted by the lingustics-only expert.

Note that partial matches can be very useful in some applications, e.g. in automatic terminology extraction or corpus pre-annotation prior to human validation. If a term is at least partly recognized the manual correction of its annotation is easy, while it might be totally overlooked otherwise. Over 98% of the manually annotated corpus terms were at least partly recognized both by our lexicon and by our grammar, which is a very good score even if many of them were non exact matches.

5 Related Work

SEJFEK is the third grammatical lexicon of Polish multi-word units built under Toposław lexicographic suite, and the first one to have been converted into a shallow grammar. Two other resources are: (i) *SAWA*[10] (Marciniak et al., 2009), a grammatical lexicon of Warsaw urban proper names containing 9,000 names of streets, squares, bus stops, monuments and other objects linked to the communication network in Warsaw, (ii) *SEJF*[11], a grammatical lexicon of Polish phraseology containing over 3,000 nominal, adjectival and adverbial compounds of the Polish general language.

A similar lexicon for Serbian (Krstev et al., 2010), containing general language compounds, was built within another lexicographic framework, *Leximir* comprising a Unitex morphological analyzer and generator module for Serbian, as well as Multiflex. This tool offers interesting facilities for automatic prediction of inflection graphs, based on rule-based mining of both the lexicon entries and the new incoming entries.

Complementary formalisms for inflectional paradigms of Polish MWUs have been presented in (Graliński et al., 2010) and (Broda et al., 2007). Like our grammar, they rely mainly on identifying the MWU's headword and checking its agreement with other components.

DuELME (Grégoire, 2010) is a lexicon of Dutch multi-word, notably verbal, expressions (MWE), which may go beyond contiguous text segments. It contains about 5,000 entries. Candidate MWEs are extracted from a corpus by pattern-based methods and filtered by a decision-tree classifier into probable true and false positives. Their variants in the corpus are analyzed in order to detect their unpredictable properties, which are definitional criteria of MWEs. Pre-selected MWE candidates are then validated and described in two steps, similar to those in SEJFEK. Firstly, the lemmas of the lexically fixed components are identified (however, unlike in SEJFEK, the morphological features of these components are stated in external parameters) and some restrictions for the non fixed components are expressed, e.g. animate object, admitted pronominalization, modal verbs going with the head component (*have* or *be*), possible adjectival modifiers, and restriction to negated use only. Secondly, the MWE is assigned a pattern. Patterns are represented as *parameterized equivalence classes* which reflect the syntactic structure of MWEs. A sample class is: *expression headed by a verb, taking a direct object consisting of a fixed determiner and a modifiable noun*, whereas an external parameter states if the object noun is in singular or in plural. Parameters allow to prevent the explosion of the number of classes. The DuELME formalism is meant to be

[10] `http://zil.ipipan.waw.pl/SAWA`
[11] `http://zil.ipipan.waw.pl/SEJF`

theory- and implementation-neutral and its applicability to a particular dependency parser has been demonstrated. We think that this description framework is very promising in that it applies to the lexical description of verbal MWEs and offers an abstract formalism, which can potentially be compiled into different parsing frameworks.

Other morphosyntactic frameworks for several European languages have been developed over the past decades. A contrastive study (Savary, 2008) shows that most of them apply one of the two complementary approaches presented in this paper: a MWU lexicon or a lexicalized local grammar. Besides Multiflex, two of these approaches, *lexc* and *FASTR*, were judged as best adapted to inflectional morphology of MWUs.

A finite-state morphology tool *lexc* (Karttunen et al., 1992; Karttunen, 1993) represents compounds by their lemmas, inflection classes and alternation rules yielding inflected forms. Like Spejd, it efficiently implements cascades of rules by a finite-state machinery. It emulates unification operators (crucial in describing agreement and government rules) and it allows the expression of various types of variations in MWUs. To the best of our knowledge, no studies report on a large-scale application of lexc to creating MWU resources.

FASTR (Jacquemin, 2001) is a shallow parser dedicated to the recognition, normalization and acquisition of compound terms, developed within a unification-based framework. FASTR's input is a corpus and an initial set of controlled complex terms that are analyzed morphologically and transformed into feature structure rules. *Metarules* can then apply to selected rules in order to model inflectional, syntactic and semantic variants of the controlled terms. As a result FASTR produces a set of links between the initial terms and occurrences of these terms and their variants in the corpus. Large coverage FASTR grammars and metagrammars have been developed for English and French terminology. Representing MWUs as fully lexicalized rules is common for FASTR and Spejd. The notable difference in Spejd is to perform both disambiguation and shallow parsing simultaneously.

Other shallow parsers have been efficiently applied to large-scale processing of Polish MWUs, notably named entities. *SProUT* (Becker et al., 2002) offers: (i) a rich grammar formalism with finite-state operators, unification and cascading, (ii) a very fast gazetteer lookup, (iii) an XML-based output in the form of typed feature structures whose type hierarchy can be defined by the user. It has been used for Polish named entity recognition (Piskorski, 2005) and annotation (Savary and Piskorski, 2011). Unlike in the Spejd grammar presented here, Polish rules in SProUT are generally less lexicalized. This fact reflects the lexical nature of named entities, in which productive structures (cf. Section 2.4) are very frequent.

Another contribution to automatic information extraction from Polish terminological texts has been presented in (Mykowiecka et al., 2009). Here again, a SProUT grammar is used, together with a medical domain ontology, a gazetteer of medical terms, and a domain-specific fine-grained grammar, in order to extract structured data from unstructured natural language mammography reports and hospital records of diabetic patients.

Conclusions and Perspectives

We have described SEJFEK, a large-coverage lexical and grammatical resource of Polish economic terminology. It consists of two alternative modules. One is a grammatical lexicon of about 11,000 terminological MWUs, where inflectional and syntactic variation, as well as nesting of terms, are described via graph-based rules. The other one is a fully lexicalized shallow grammar of a roughly equal number of rules, obtained by an automatic conversion

of the lexicon, and partly manually validated.

SEJFEK is the first NLP-oriented resource for Polish economic terminology and one of the first resources of this kind for Slavic languages. It is freely available[12] under the Creative Commons BY-SA license[13]. It might be used in automatic term extraction, document classification, domain-specific information extraction or question answering, or any application where a reliable inflection-aware identification and conflation of terms and their variants is crucial. As a means of term normalization it might also be useful in professional writing support software, such as *Acrolinx*[14], or in computer-assisted translation tools which allow users to import external terminology, e.g. *SLD Trados Multiterm Desktop*[15].

Both resources show a good and largely comparable coverage, which demonstrates the complementarity of a lexicon and a fully lexicalized grammar. The evaluation results, obtained on a 221,000-token manually annotated economic corpus, show the MWU-per-WMU correctness of over 41% and the token-per-token correctness of more than 68%. About 98% of all corpus terms are at least partly recognized by both the lexicon and the grammar. The main advantage of the lexicon-to-grammar conversion lies in the fact that the entire lexico-syntactic knowledge contained in a lexicon entry can be explicitly expressed in the structured output of the grammar. This result contributes to a better lexicon-grammar interface as far as the treatment of MWUs is concerned.

Since the lexicon-to-grammar conversion does not exploit the internal semantics of lexicon's inflection graphs, it fails to account for some syntactic variants of terms (word order changes, ellipses, acronyms optional inflection, etc.). However its strength lies in the fact that it can operate on roughly annotated input data. Thus, it might be used reversely: (i) it might yield approximate grammar rules in order to match text occurrences of a new term, (ii) these occurrences might help match or develop graphs in Toposław for new lexicon entries.

Other perspectives include: (i) completing Morfeusz' lexicon in order to cover all components appearing in our resource, notably foreign proper names, (ii) editing a proofread version of the resource resulting from the Morfeusz completion and from an analysis of conversion errors, (iii) involving a second annotator, expert in economy or in translation of economic texts, on order to increase the corpus quality, (iv) completing the grammar by partly non-lexicalized rules covering productive patterns, as those mentioned in Sec. 2.4, (v) designing a standard LMF[16] exchange format (possibly both lexicon- and grammar-compatible), (vi) a better automation of graph matching in Toposław inspired by (Krstev et al., 2006), (vii) exploiting the internal structure of graphs during conversion in case a higher-precision grammar is needed.

Acknowledgments

This work has been carried out within two projects: (i) *Nekst*[17], funded by the European Regional Development Fund and the Polish Ministry of Science and Higher Education, (ii) CESAR[18] - a European project (CIP-ICT-PSP-271022), part of META-NET.

[12] http://zil.ipipan.waw.pl/SEJFEK
[13] http://creativecommons.org/licenses/by-sa/3.0/
[14] http://www.acrolinx.com/terminology_support.html
[15] http://www.translationzone.com/en/translator-products/sdlmultitermdesktop/
[16] http://www.lexicalmarkupframework.org/
[17] http://www.ipipan.waw.pl/nekst/
[18] http://www.meta-net.eu/projects/cesar

-

References

Alex, B., Haddow, B., and Grover, C. (2007). Recognising nested named entities in biomedical text. In *BioNLP '07: Proceedings of the Workshop on BioNLP 2007*, pages 65–72, Morristown, NJ, USA. Association for Computational Linguistics.

Becker, M., Drożdżyński, W., Krieger, H.-U., Piskorski, J., Schäfer, U., and Xu, F. (2002). SProUT - Shallow Processing with Typed Feature Structures and Unification. In *Proceedings of ICON 2002, Mumbay, India.*

Black, J. (2008). *Słownik ekonomii.* Wydawnictwo Naukowe PWN, Warszawa.

Broda, B., Derwojedowa, M., and Piasecki, M. (2007). Recognition of structured collocations in an inflective language. In *Proceedings of the International Multiconference on Computer Science and Information Technology — 2nd International Symposium Advances in Artificial Intelligence and Applications (AAIA'07)*, pages 237–246.

Chow, G. C. (1995). *Ekonometria.* Wydawnictwo Naukowe PWN, Kraków.

Daille, B. (1996). Study and implementation of combined techniques for automatic extraction of terminology. In Klavans, J. L. and Resnik, P., editors, *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, pages 49–66. MIT Press, Cambridge.

Finkel, J. R. and Manning, C. D. (2009). Nested Named Entity Recognition. In *Proceedings of EMNLP-2009*, Singapore.

Graliński, F., Savary, A., Czerepowicka, M., and Makowiecki, F. (2010). Computational Lexicography of Multi-Word Units: How Efficient Can It Be? In *Proceedings of the COLING-MWE'10 Workshop, Beijing, China.*

Grégoire, N. (2010). DuELME: a Dutch electronic lexicon of multiword expressions. *Language Resources and Evaluation*, 44(1-2).

Głuchowski, J. and Szambelańczyk, J., editors (1999). *Bankowość. Podręcznik dla studentów.* Wydawnictwo Wyższej Szkoły Bankowej, Poznań.

Jacquemin, C. (2001). *Spotting and Discovering Terms through Natural Language Processing.* MIT Press.

Karttunen, L. (1993). Finite-State Lexicon Compiler. Technical Report ISTL-NLTT2993-04-02, Xerox PARC.

Karttunen, L., Kaplan, R. M., and Zaenen, A. (1992). Two-Level Morphology with Composition. In *Proceedings of the 14 th International Conference on Computational Linguistics (COLING'92), Nantes*, pages 141–148.

Koeva, S. (2007). Multi-word term extraction for bulgarian. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing*, pages 59–66, Prague, Czech Republic. Association for Computational Linguistics.

Krstev, C., Stankovic, R., Obradovic, I., Vitas, D., and Utvic, M. (2010). Automatic construction of a morphological dictionary of multi-word units. In *Proceedings of IceTAL 2010*, volume 6233 of *Lecture Notes in Computer Science*, pages 226–237.

Krstev, C., Stanković, R., Vitas, D., and Obradović, I. (2006). WS4LR: A Workstation for Lexical Resources. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy*, pages 1692–1697.

Kuciński, K., editor (2009). *Geografia ekonomiczna*. Szkoła Główna Handlowa, Kraków.

Marciniak, M., Rabiega-Wiśniewska, J., Savary, A., Woliński, M., and Heliasz, C. (2009). Constructing an Electronic Dictionary of Polish Urban Proper Names. In *Recent Advances in Intelligent Information Systems*, pages 233–246. Exit.

Marciniak, M., Savary, A., Sikora, P., and Woliński, M. (2011). Toposław - a lexicographic framework for multi-word units. *Lecture Notes in Computer Science*, 6562:139–150. Springer.

Michalski, E. (2003). *Marketing. Podręcznik akademicki*. Wydawnictwo Naukowe PWN, Warszawa.

Michoń, F., editor (1991). *Ekonomika pracy: zarys problematyki i metod*. Państwowe Wydawnictwo Naukowe, Kraków.

Mykowiecka, A., Marciniak, M., and Kupść, A. (2009). Rule-based information extraction from patients' clinical data. *Journal of Biomedical Informatics*, 42(5):923–936.

Paumier, S. (2008). Unitex 2.1 User Manual.

Piskorski, J. (2005). Named-Entity Recognition for Polish with SProUT. In *LNCS Vol 3490: Proceedings of IMTCI 2004, Warsaw, Poland.*

Przepiórkowski, A. (2008). *Formalizm ♠*, chapter 7. Akademicka Oficyna Wydawnicza EXIT, Warsaw.

Przepiórkowski, A. and Buczyński, A. (2007). ♠: Shallow parsing and disambiguation engine. In *Proceedings of the 3rd Language & Technology Conference*, Poznań.

Rynarzewski, T. and Zielińska-Głębocka, A. (2006). *Międzynarodowe stosunki gospodarcze. Teoria wymiany i polityki handlu międzynarodowego*. Wydawnictwo Naukowe PWN, Warszawa.

Rzewuska, M., Gałkiewicz, A., and Falkenberg, J., editors (2001). *Ekonomia — finanse — pieniądz: glosariusz angielski–francuski–niemiecki–polski*. Urząd Komitetu Integracji Europejskiej, Warszawa.

Samuelson, A. and Nordhaus, W. D. (1998). *Ekonomia*, volume 2. Wydawnictwo Naukowe PWN, Warszawa.

Samuelson, A. and Nordhaus, W. D. (2003). *Ekonomia*, volume 1. Wydawnictwo Naukowe PWN, Warszawa.

Savary, A. (2008). Computational Inflection of Multi-Word Units. A contrastive study of lexical approaches. *Linguistic Issues in Language Technology*, 1(2):1–53.

Savary, A. (2009). Multiflex: a Multilingual Finite-State Tool for Multi-Word Units. *Lecture Notes in Computer Science*, 5642:237–240.

Savary, A. and Piskorski, J. (2011). Language Resources for Named Entity Annotation in the National Corpus of Polish. *Control and Cybernetics*, 40(2):361–391.

Savary, A., Rabiega-Wiśniewska, J., and Woliński, M. (2009). Inflection of Polish Multi-Word Proper Names with Morfeusz and Multiflex. *Lecture Notes in Computer Science*, 5070:111–141.

Smadja, F. (1993). Xtract : An overview. *Computer and the Humanities*, 26:399–413.

Smullen, J. and Hand, N., editors (2008). *Słownik finansów i bankowości*. Wydawnictwo Naukowe PWN, Warszawa.

Treder, H. (2005). *Podstawy handlu zagranicznego*. Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk.

Wernik, A. (2007). *Finanse publiczne. Cele, struktury, uwarunkowania*. Polskie Wydawnictwo Ekonomiczne, Warszawa.

Wilcock, G. (2009). *Introduction to Linguistic Annotation and Text Analytics*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Zaborowski, B. (2012). *Spejd 1.3.6 - User manual*.

Śnieżek, E., editor (2004). *Wprowadzenie do rachunkowości — podręcznik z przykładami, zadaniami i testami*. Oficyna Ekonomiczna, Kraków.