Université François Rabelais – Tours, antenne de Blois

Laboratoire d'Informatique

(UPRES EA n°2101)

ProlexFeeder — Populating a Multilingual Ontology of Proper Names from Open Sources

Auteurs:
Agata SAVARY
Leszek MANICKI
Małgorzata BARON

Address:
Laboratoire d'Informatique
Ecole Polytechnique –
Département Informatique
64 avenue Jean Portalis
37200 TOURS – FRANCE
Tél.: 02.47.36.14.14
Fax: 02.47.36.14.36
email:
agata.savary@univ-tours.fr

5 April 2013
Technical Report n°306, 38 pages
Extended version of a paper
submitted to the *Journal of Language Modelling*

Journal of Language Modelling Vol 0, No 0 (1972)
ISSN 666-666-666, pages 1–38

ProlexFeeder — Populating a Multilingual Ontology of Proper Names from Open Sources

Agata Savary[1], Leszek Manicki[2,3], and Małgorzata Baron[1,2]

[1] Université François Rabelais Tours, France
[2] Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland
[3] Poleng, Poznań, Poland

Abstract

Proper names and, more generaly, named entities, carry a particularly rich semantic load in each natural language text and their central role in natural language processing (NLP) applications is unquestionable. They are still underrepresented in lexicons, annotated corpora, and other resources dedicated to text processing. One of the main challenges is both the prevalence and the dynamicity of proper names. At the same time, large and regularly updated knowledge sources containing partially structured data, such as Wikipedia or GeoNames, are publicly available and contain large numbers of proper names. We present a method for a semi-automatic enrichment of Prolexbase, an existing multilingual ontology of proper names dedicated to natural language processing, with data extracted from these open sources in three languages: Polish, English and French. Fine-grained data extraction and integration procedures allow the user to enrich previous contents of Prolexbase with new incoming data. All data are manually validated and available under an open licence.

Keywords: proper names, named entities, multilingual ontology population, Prolexbase, Wikipedia, GeoNames, Translatica

1 Introduction

Proper names and, more generally, named entities, carry a particularly rich semantic load in each natural language text since they refer to persons, places, objects, events and other entities crucial for its understanding. Their central role in natural language processing (NLP) applications is unquestionable but they are still underrepresented in lexicons, annotated corpora, and other resources dedicated to text processing. One of the main challenges is both the prevalence and the dynamicity of proper names. New names are constantly created for new institutions, products and works. New individuals or groups of people are brought into focus and their names enter common vocabularies.

At the same time, large knowledge sources become publicly available, and some of them are constantly developed and updated by a collaborative effort of large numbers of users, Wikipedia being the most prominent example. The data contained in these sources are partly structured, which increases their usability in automatic text processing.

In this paper our starting point is Prolexbase (Krstev *et al.*, 2005; Tran and Maurel, 2006; Maurel, 2008), an open multilingual knowledge base dedicated to the representation of proper names for NLP applications. Prolexbase initially contained mainly French proper names, even if its model supports multilingualism. In order to extend its coverage of other languages we created *ProlexFeeder*, a tool meant for a semi-automatic population of Prolexbase from open sources, notably Wikipedia and GeoNames.

Figure 1 shows the dataflow in our Prolexbase population process. Three main data sources are: (i) Polish, English and French Wikipedia, (ii) Polish names in GeoNames, (iii) Polish inflection resources in Translatica, a machine translation software. Automatically selected relevant classes in Wikipedia and in GeoNames are manually mapped on Prolexbase typology. The data belonging to the mapped classes are automatically extracted and their popularity (or frequency) is estimated. Inflection rules are used to automatically predict inflected forms of both simple and multi-word entries from Wikipedia. The resulting set of candidate names is fed to ProlexFeeder, which integrates them with Prolexbase in two steps. Firstly, it is automatically checked if the entity represented by a candidate entry is already present in Prolexbase. Secondly, the entry, together with its translations, variants, relations and inflected forms is manually validated by an expert lexicographer.

The motivation behind Prolexbase is not to represent as many available names as possible, like in the case of other large automatically constructed ontologies such as YAGO (Suchanek *et al.*, 2007) or DBpedia (Mendes *et al.*, 2012). We aim instead at a high quality, i.e. **manually validated**, **incremental** resource **dedicated to NLP**. This implies:

- A rather **labour-intensive** activity, thus a **reduced scope** of the result. This requires a definition of appropriate selection criteria that allow us to retain only the most relevant, popular and stable names. In this paper we exploit criteria based on: (i) the popularity of the corresponding Wikipedia articles, (ii) systematic lists of some major categories found in GeoNames.
- Thorough **data integration** techniques allowing us to avoid duplication of data during an enrichment process (as opposed to extraction from scratch) in which previously validated data can be merged or completed with new incoming data.
- **NLP-targeted features**, particularly with respect to **highly inflected languages** such as Polish, non-existent in traditional ontologies. Prolexbase was designed with such languages in mind, notably Serbian (Krstev *et al.*, 2005), which belongs, like Polish, to the family of Slavic languages. This allows us to account for rich word formation, variation and inflection processes within the same model.

Potential applications of such a limited-scope but high-quality resource are

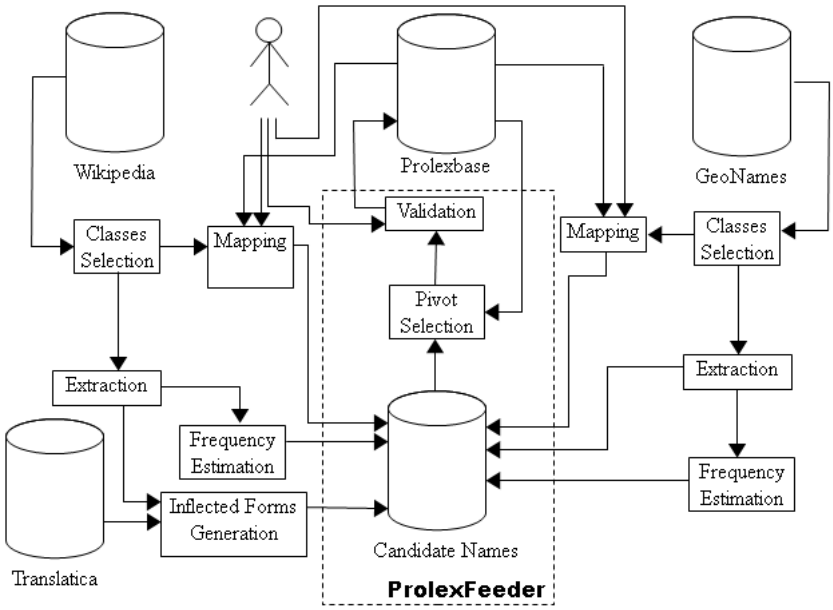

FIGURE 1: Data flow in Prolexbase population via ProlexFeeder.

manifold and include: (i) uni- and multilingual named entity recognition (NER) (Richman and Schone, 2008), (ii) word sense disambiguation (Fernando and Stevenson, 2012), (iii) machine translation (Graliński *et al.*, 2009), (iv) text classification (Kumaran and Allan, 2004), (vi) uni- or cross-lingual question answering (Ferrández *et al.*, 2007), (vii) proper name normalization (Jijkoun *et al.*, 2008) and others.

Due to resources such as Prolexbase, novel viewpoints might be developed in particular on the long-established tradition of the NER activity. Namely, most NER systems – existing and newly created for increasingly many languages – have a relatively good capacity of spotting and categorizing names which appear frequently in training corpora. However, it becomes increasingly difficult to attain the cutting edge, i.e. to improve the NER performance in the tail of the frequency distribution. In order to boost the research dedicated to these scarcity problems, we think that NER evaluation methodology might be redesigned so as to make a bigger impact on rare and unstable named entities. A possible approach would be here to consider a high quality, NLP-oriented and manually validated resource of the most popular names, such as Prolexbase, as a publicly available baseline. All names outside of this baseline would then be given the crucial role in the evaluation of the added value of a NER system's outcome.

This paper is organized as follows. Section 2 summarizes the major features of Prolexbase and of the input data sources relevant to the population process. Section 3 describes data integration issues. In Section 4 we show extracts of the human validation interface and Section 5 is dedicated to evaluation of the population process. Section 6 contains a detailed discussion of related work. The last

section summarizes the results and mentions some perspectives for future work.

2 Input Knowledge Sources

2.1 Prolexbase

Prolexbase (Krstev *et al.*, 2005; Tran and Maurel, 2006; Maurel, 2008) offers a fine-grained multilingual model of proper names whose specificity is to be both concept-oriented and lexeme-oriented. Namely, it comprises a language-independent ontology of concepts referred to by proper names, as well as detailed lexical modules for proper names in several languages (French, English, Polish and Serbian being the best covered ones). Prolexbase is structured in four levels for which a set of relations is defined.

The **metaconceptual level** defines a two-level typology of four **supertypes** and 34 **types**, cf. Agafonov *et al.* (2006):

1. *Anthroponym* is the supertype for individuals — *celebrity*, *first name*, *patronymic*, *pseudo-anthroponym* — and collectives — *dynasty*, *ethnonym*, *association*, *ensemble*, *firm*, *institution*, and *organization*.
2. *Toponym* comprises territories — *country*, *region*, *supranational* — and other locations — *astronym*, *building*, *city*, *geonym*, *hydronym*, and *way*.
3. *Ergonym* includes *object*, *product*, *thought*, *vessel*, and *work*.
4. *Pragmonym* contains — *disaster*, *event*, *feast*, *history*, and *meteorology*.

Some types have secondary supertypes, e.g. a city is not only a toponym but also an anthroponym and a pragmonym. The metaconceptual level contains also the **existence** feature which allows to state if a proper name referent has really existed (*historical*), has been invented (*fictitious*) or whether its existence depends on religious convictions (*religious*).

The originality of the **conceptual level** is twofold. Firstly, proper names designate concepts (called **conceptual proper names**), instead of being just instances of concepts, as in the state-of-the-art approaches discussed in Section 6. Secondly, these concepts, called **pivots**, include not only objects referred to by proper names, but also points of view on these objects: *diachronic* (depending on time), *diaphasic* (depending on the usage purpose) and *diastratic* (depending on sociocultural stratification). For instance, although *Alexander VI* and *Rodrigo Borgia* refer to the same person, they get two different pivots since they represent two different points of view on this person. Each pivot is represented by a unique interlingual identification number allowing to connect proper names that represent the same concepts in different languages. Pivots are linked by three language-independent relations. **Synonymy** holds between two pivots designating the same referent from different points of view (*Alexander VI* and *Rodrigo Borgia*). **Meronymy** is the classical relation of inclusion between the meronym (*Samuel Beckett*) and the holonym (*Ireland*). **Accessibility** means that one referent is accessible through another one, generally better known (Tran and Maurel, 2006). The accessibility **subject file** with 12 values (*relative*, *capital*, *leader*, *founder*, *follower*, *creator*, *manager*, *tenant*, *heir*, *headquarters*,

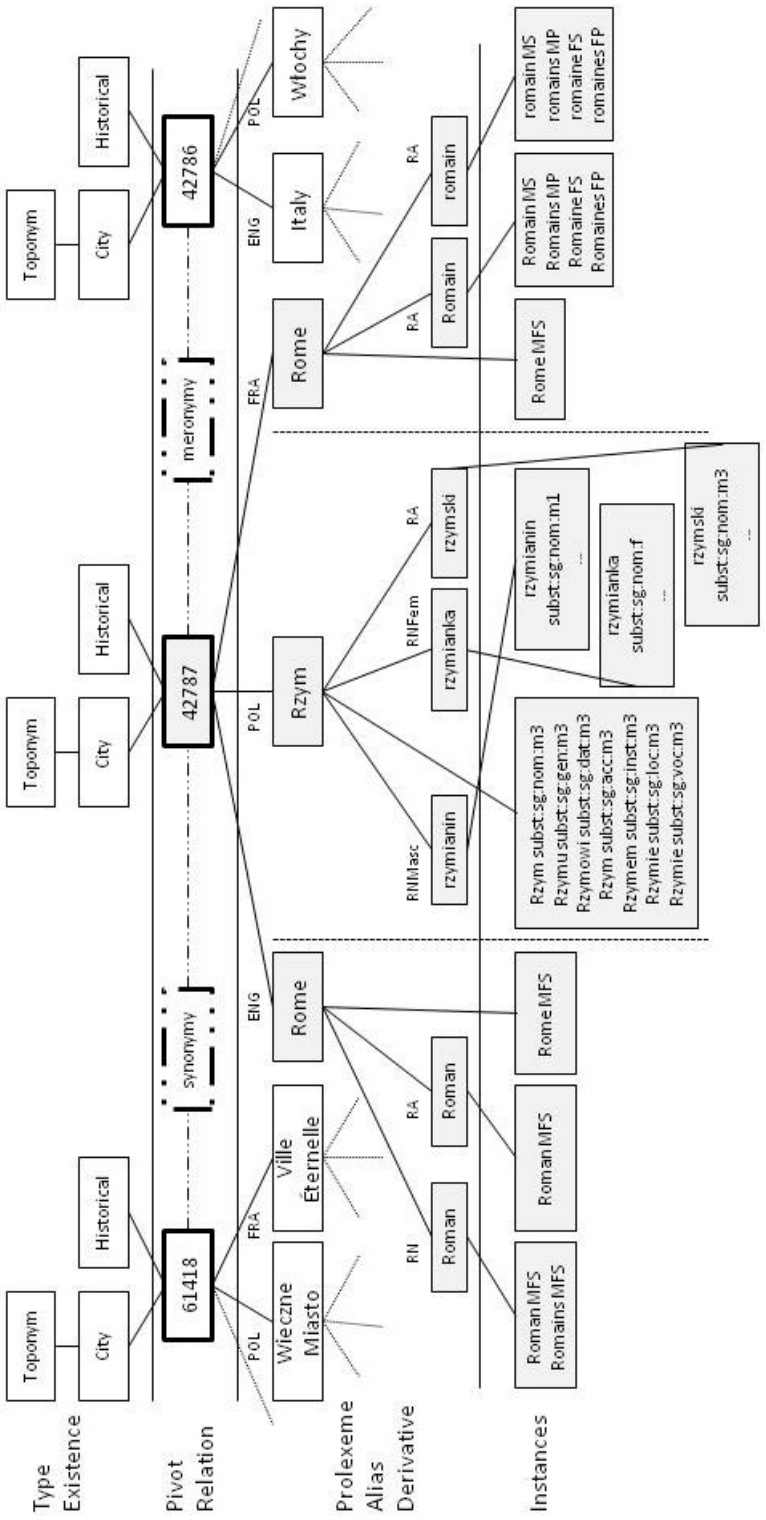

FIGURE 2: Extract of the intended contents of Prolexbase with four levels and three prolexemes.

rival, and *companion*) informs us about how/why the two pivots are linked (*The Magic Flute* is accessible from *Mozart* as *creator*).

The **linguistic level** contains **prolexemes**, i.e. the lexical representations of pivots in a given language. For instance, pivot 42786 is linked to the prolexeme *Italy* in English, *Italie* in French and *Włochy* in Polish. There is a 1 : 1 relation between pivots and prolexemes within a language, thus homonyms (*Washington* as a celebrity, a city and a region) are represented by different prolexemes. A prolexeme can have language-dependent variations: **aliases** (abbreviations, acronyms, spelling variants, transcription variants, etc.) and **derivatives** (relational nouns, relational adjectives, prefixes, inhabitant names, etc.). The language-dependent relations defined at this level are: **collocation** (*en France*, *au Portugal*), **classifying context** (the *Vistula river*), **accessibility context** (*Paris* — the *capital* of *France*), *eponymy* (lexeme or expression morphologically linked to but semantically distant from a proper name: e.g. *a French leave*), **frequency** (*commonly used*, *infrequently used* or *rarely used*), **sort** (*George Washington* should be lexicographically sorted with respect to the family name at first), **language** (association of each prolexeme to one language) and **phonetics** (pronunciation of a foreign lexeme in a given language).

The **level of instances**[1] contains inflected forms of prolexemes, aliases and derivatives, together with their morphological or morphosyntactic tags. These forms can either be materialized within Prolexbase itself or be represented by links to external morphological models and resources.

Fig. 2, inspired by Krstev *et al.* (2005), shows an extract of the intended contents of Prolexbase containing the vicinity of the prolexeme *Rzym* 'Rome', in particular its pivot, stylistic synonym, meronym, derivatives, and instances.

Prolexbase is a model dedicated to NLP applications. Named entity recognition and categorisation can be benefit from Prolexbase morphosyntactic description via aliases and instances, as well as from its rather rich typology (Maurel *et al.*, 2011). Coreference annotation may use instances, aliases, derivatives and inter-pivot relations in order to conflate different surface names referring to the same object. In machine translation of proper names, the original idea of a conceptual proper name being a pair of a referred object and a point of view on this object allow the user application to provide the most appropriate equivalent (rather than just any equivalent) for a name in other languages. For some names, popular in one language but unknown in others, relations like classifying context and accessibility enable explanation-based translations (e.g. *Hanna Gronkiewicz Walz* ⇒ *Hanna Gronkiewicz-Walz, the president of Warsaw*).

In order to adapt Prolexbase to being populated with Wikipedia data in an automated way several minor changes in the original Prolexbase structure have been made. Notably, the **Wikipedia link** attribute has been added to the description of prolexemes in every language. Furthermore, since intense searches of prolexemes,

[1] Note that Prolexbase terminology is non-standard with respect to WordNet (Miller, 1995). Notably, in Prolexbase hypernyms of entities referred to by proper names are *meta-concepts*, entities are *concepts* (represented by pivot identifiers), and the inflected forms of names are called *instances*. In WordNet, hypernyms of entities are *concepts* while the surface forms of the entities themselves, are called *instances*. See also Section 6 for a discussion on the instance-to-concept mapping which we perform, as opposed to the concept-to-concept mapping standard in the related work.

aliases and instances are frequently performed by ProlexFeeder, indices have been created on appropriate data.

2.2 Wikipedia

Wikipedia is a constantly growing project grouping a set of freely-available on-line encyclopaedia initiatives run and filled with content by volunteer authors. It is based on the **wiki**[2] concept, invented by Ward Cunningham in 1994, of software that allows any user to be involved in quick collaborative creation and modification of website content using a simplified markup language. It also focuses on associations between the user-created data, which often results in a frequent use of hyperlinks.

Wikipedia is run by the MediaWiki software and maintained by the non-commercial Wikimedia Foundation, which also supports several other projects based on the wiki concept. At the beginning of October 2012 there were 275 active language (or dialect) versions of Wikipedia consisting of the total number of 23,296,142 articles[3]. Four Wikipedias comprise more than a million articles: English (over 4 million articles), German, French and Dutch language versions. In this paper we focus mainly on the Polish version, which is the sixth largest Wikipedia project with its 900,000 articles.

The Wikipedia content is available under the open GFDL license, permitting the redistribution of data. The main access mode is the on-line access to Wikipedia's own websites or mirrors. Moreover, the Wikimedia Foundation regularly releases updated dumps of each language version of Wikipedia and of related projects. Various release packages differ in the range of content covered in the dump. In our project we used the dump containing all Polish articles of the current release available at the beginning of 2011, with all metadata and internal MediaWiki software data excluded. The data extraction process described in Section 3.1.1 may be iteratively repeated using a newer Wikipedia dump release in order to add new entries to Prolexbase and complete the existing ones.

Each **article** describing a Wikipedia entry is written in the MediaWiki markup language, and is rendered by the wiki software as an HTML webpage. Articles describing the same topic in different languages can be connected by **interwiki** links. We used this interconnection shorthand feature to automatically extract translations for titles of Polish articles.

There are several special classes of Wikipedia articles, notably **categories**. Articles may be added to a category using simple MediaWiki mark-up tags. Categories are automatically indexed and the list of all articles from the given category is automatically generated by the MediaWiki software. The list of all categories the article belongs to is printed at the end of its webpage.

Another useful facility provided by the MediaWiki software is an **infobox**, i.e. a table summarizing an article's most important data in a structured form. Different infobox templates exist for different types of entries. For instance, an infobox dedicated to a person contains a field for his/her birth date, nationality, activity domain, etc. while an infobox of a river shows its origin, mouth, length, etc.

[2] The term *wiki* comes from a Hawaiian word meaning *fast* or *quick*.

[3] `http://s23.org/wikistats/wikipedias_html.php?sort=good_desc`

Thus, the infobox templates can be used as a secondary typology for the entries containing infoboxes.

Both Wikipedia categories and infoboxes are language-specific and user-defined. No mechanism is provided for ensuring a compatibility of category hierarchies in different languages. As a result, a Polish entry and its English or French equivalent may be assigned to non-equivalent categories or incompatible category hierarchies. Moreover, categories are often used by Wikipedia users to group related articles rather than to create a hierarchical structure of data. As a result some categories may include both individual entities and general domain-related terms. For instance, the category *powiaty* 'counties' in Polish Wikipedia contains the list of Polish counties but also terminological items such as *powiat grodzki* 'city county' (a county type in the current Polish administrative division system) or *powiaty i gminy o identycznych nazwach* 'Polish homonymous counties and communes' (containing the list of homonymous Polish administrative division units). Conversely, infoboxes are usually added to articles that only cover individual entities, not general domain-related terms. For this reason we used infobox templates as the main criteria for extracting and classifying proper names from Wikipedia, as described in Section 3.1.1.

Like categories, **redirects** belong to special classes of Wikipedia articles. They allow to automatically access an article whose title is not identical with the query. Redirects may be used for various purposes including those in examples (1)–(8).

(1) Main Polish entry: *Snåsa* (Norwegian city)
Redirects: *Snaasa*, *Snasa*, *Snåase*
Redirect type: orthography variants

(2) Main English entry: *Boris Yeltsin*
Redirects: *Boris Elcin*, *Boris Eltsin*, *Boris Jel'cin*, *Boris Jelcin*, *Boris Jeltsin*, *Boris Jelzin*
Redirect type: transcription variants

(3) Main Polish entry: *Wielka Brytania* 'Great Britain'
Redirects: *Zjednoczone Królestwo* 'United Kingdom', *Zjednoczone Królestwo Wielkiej Brytanii i Irlandii Północnej* 'United Kingdom of Great Britain and Northern Ireland'
Redirect type: official names and their elliptical variants

(4) Main Polish entry: *Warszawska Brygada Pancerno-Motorowa* 'Warsaw Armoured Motorized Brigade'
Redirects: *WBPM*, *WBPanc-Mot*
Redirect type: acronyms and abbreviations

(5) Main Polish entry: *Plac Powstańców Warszawy w Warszawie* 'Warsaw Uprising Square in Warsaw'
Redirects: *Plac Napoleona* 'Napoleon Square', *Plac Warecki* 'Warka Square'
Redirect type: different names of the same entities, notably in different historical periods

(6) Main English entry: *Marilyn Monroe*
Redirects: *Norma Jeane Mortenson*, *Norma Jeane Baker*
Redirect type: variants of people's names, including full family names, pre-

vious names, pseudonyms, etc.

(7) Main English entry: *Poland*
Redirects: *Ploand*, *Poleand*, *Polland*, *Polnd*
Redirect type: common spelling errors

(8) Main English entry: *Sierra Blanca* (settlement in Texas)
Redirects: *Sierra Blanca (TX)*, *Sierra Blanca, TX*
Redirect type: names with extra disambiguating data

Since some of the redirect types in examples (1) through (8) correspond to the Prolexbase ideas of aliases and synonyms they were extracted automatically and validated manually, as described in Sections 3.1.1 and 4.

2.3 GeoNames

GeoNames is a database of geographical names collected from various publicly available and official sources such as American National Geospatial-Intelligence Agency (NGA), U.S. Geological Survey Geographic Names Information System or British Ordnance Survey. The database contains over 10 million records related to over 8 million unique features. It stores toponym names in different languages but also some encyclopaedic and statistical data such as elevation, population, latitude and longitude. Information on administrative subdivision is also provided for numerous entries. All the data are freely available under the Creative Commons Attribution license[4], both through the GeoNames web interface and through numerous programming libraries (APIs). As GeoNames exploits heterogeneous sources and the quality of its contents may vary, a wiki-like interface is provided for users in order to correct and expand the data.

GeoNames entries are categorized into 9 main classes which in turn divide into 645 highly fine-grained subcategories.[5] Each category is labelled with a letter while the subcategories are labelled with *feature codes*. Refer to Section 3.2.2 as well as Tab. 2 and Tab. 3 for the codes of GeoNames main categories and selected subcategories.

2.4 Translatica

As described in Section 2.1, Prolexbase entries in any language are supposed to be supplied with their inflected forms called *instances*. Neither GeoNames, nor Wikipedia, which is an encyclopaedia rather than a dictionary, contain explicit inflection or grammatical data. Due to the limited inflection system of English and French proper names we do not automatically generate inflected forms of entries in these languages. Polish, however, contrary to English and French, has a rich inflection system and instances have to be suggested automatically if the human validation is to be efficient. We use the following inflection routine developed for Translatica (Jassem, 2004), a Polish-centred machine translation system:

[4] `http://creativecommons.org/licenses/by/3.0/`
[5] `http://www.geonames.org/export/codes.html`

1. Single-word lemmas are looked up in a dictionary of inflected forms. If a lemma appears in the dictionary the list of its inflected forms is retrieved. Otherwise the inflected forms are predicted by analogy to known lemmas.
2. Multi-word proper names are automatically assigned intentional inflection paradigms expressed in the POLENG formalism (Author et al. 2010). Its implementation allows to automatically generate the extensional list of inflected forms.

While the former step seems rather straightforward, the latter one requires additional explanation. The POLENG formalism was designed both for recognition and for generation of multi-word units (MWUs) in Translatica. Inflection paradigms of MWUs are expressed as compact strings.

```
R:4  [ światowy_P:u  R:u!       0       ]
       Światowa      Organizacja  Zdrowia
```

FIGURE 3: Inflection paradigm of *Światowa Organizacja Zdrowia* 'World Health Organization' in Translatica.

Fig. 3 shows a sample inflection paradigm for a Polish organization name. The left-most capital letter stands for the part-of-speech of the whole MWU (*Rzeczownik* 'noun'). Its gender is feminine (*4*). Its components are described in square brackets. Each component which inflects within the MWU is morphologically identified by its lemma, part-of-speech and additional flags (if any). For instance the tag `światowy_P:u` describes the first component *Światowa* '[worldwide]$_{feminine}$', which has the lemma *światowy* '[worldwide]$_{masculine}$' and is an adjective (*Przymiotnik*) in uppercase (`u`).

If the lemma of a component taken as an individual word is the same (up to letter case) as in the MWU's lemma it can be omitted. For instance `R:u` refers to *Organizacja* 'organization' whose lemma is *organizacja*. If a component never varies within the MWU it is referred to by a *0*. That is the case here of *Zdrowia* 'Health$_{genitive}$, which always remains in genitive: *Światową Organizacją Zdrowia*, *Światowej Organizacji Zdrowia*, etc.

It is implicitly assumed that all inflected components agree in case, number and gender with the head component, marked with an '!'. Here, the modifier *Światowa* agrees with the head noun *Organizacja*.

Most Polish entries extracted from Wikipedia were inflected by the above procedure. Namely, for over 260,000 extracted entries almost 2 million instances have been collected. We used the *Morfologik* dictionary[6] as a source of inflected forms both for single entries and for MWU components. All Polish instances obtained in this way were further manually validated and corrected before their addition to Prolexbase (cf Section 4).

[6] `http://morfologik.blogspot.com`

3 Data Integration

3.1 Data Selection

Wikipedia and GeoNames were used as the main sources of new entries for Prolexbase enrichment. In this section we describe the process of extracting possibly most relevant, complete and structured data from both sources.

3.1.1 Data Selection from Wikipedia

Since Wikipedia is a general-purpose encyclopaedia, the first challenge was to select only those Wikipedia articles whose titles represent proper names. Initially, Wikipedia categories seemed to provide natural selection criteria. Some previous attempts, such as (Toral *et al.*, 2008), are based, indeed, on mapping WordNet synsets onto Wikipedia categories and on applying capitalisation rules for retaining only virtual proper names form a set of entries. However the high number of Wikipedia categories (1,073 low-level in Polish, 73,149 in total) and their heterogeneous nature explained in Section 2.2 made us turn to primarily using **infoboxes**, similarly to DBpedia (Bizer *et al.*, 2009).

We extracted the list of all infobox templates used in Polish Wikipedia and manually selected those which seemed related to proper names. As a result we obtained 340 relevant templates. We extracted all Polish entries containing infoboxes built upon these templates. Each entry was assigned a class based on the name of the corresponding infobox template. English and French translations of Polish entities (if any) were extracted via interwiki links. Thus, we obtained a trilingual list of classified named entities, henceforth called *initWikiList*.

The Polish version of Wikipedia, unlike e.g. the English version, contains rather few infobox templates referring to people. Even if several specific classes, like *żołnierz* 'soldier', *piłkarz* 'football player' or *polityk* 'politician' do exist, the major part of people-related articles contain a *biogram* 'personal data' infobox, consisting only of basic personal data (date of birth, nationality, etc.). The *initWikiList* contained numerous Polish entries with an infobox of the *biogram* class. We noticed that such entries often belong to fine-grained Wikipedia **categories**, e.g. *niemieccy kopozytorzy baroku* 'German Baroque composers'. These categories turned out to be rather homogeneous in terms of including only actual named entities, and not general domain-related terms (cf Section 2.2). Moreover, many articles belonging to these categories had no infobox attached.

This observation led us to extending the coverage of the extraction process. We collected the list of 676 person-related categories containing entries from *initWikiList*. Then we expanded *initWikiList* with all entries from these categories that did not contain an infobox. Each entry from the resulting trilingual list was assigned: (i) its Wikipedia URLs in Polish, English and French (if any) (ii) its *Wikipedia class*, i.e. its Polish infobox class (if its article contained an infobox) or its Polish category (if the entry belonged to a person-related Wikipedia category). After filtering out some evident errors we obtained the final list of candidate proper names and their associated data to be added to Prolexbase. The list contained 262,124 Polish entries with 255,835 English and 139,770 French translations.

As mentioned in Section 2.2, Wikipedia **redirects** may be valuable sources of aliases and synonyms for the retrieved entries but they are of heterogeneous nature. Types exemplified by (1)–(4) represent aliases in terms of Prolexbase. Types (5) and (6) correspond to diachronic and diastratic synonyms, respectively. Types (7)–(8) are irrelevant. The redirect type is hard to distinguish automatically. We could automatically remove only redirects of type (8), as well as those pointing at article subsections rather than articles themselves. The elimination of type (7), as well as the distinction between virtual aliases and synonyms had to be left for further manual validation stage (cf. Section 4). The final collection result contained 33,530 redirects to Polish, 297,377 to English, and 92,351 to French Wikipedia articles.

3.1.2 Data Selection from GeoNames

As the amount of data stored in GeoNames is enormous it is hardly feasible to validate them manually before adding to Prolexbase. Thus, it was necessary to select a well-defined subset of these data. We have used only the country names[7], all Polish names[8], as well as alternate names[9]. We have examined several category-dependent selection criteria based on numerical data accessible in GeoNames, such as the height of a mountain or the population of a city. Such criteria proved hard to apply in a general case: some well known mountains or cities are low or have few inhabitants. We finally decided to treat GeoNames as complementary to Wikipedia as far as the selection criteria are concerned. Namely, Wikipedia entries were sorted by their *frequency* value based on the popularity of the corresponding articles in Wikipedia, as discussed in Section 3.3. Conversely, GeoNames was used as a source of systematic lists of names belonging to some major categories. Up till now the following GeoNames categories have been selected: (i) all countries and their capitals, (ii) all first-order (*województwo*) and second-order (*gmina*) administrative division units in Poland and their chief towns, (iii) all first-order administrative division units in other European countries and their chief towns. Other GeoNames entries were extracted only if they referred to entities located in Poland. The total number of entries selected from GeoNames according to these criteria was equal to 42,376.

3.2 Ontology Mapping

Merging different ontologies into a common structure is a well known problem, as discussed in Section 6. In most approaches, the aim is to propose a unified framework in which one ontlogy is mapped on another one and the granularity of both can be fully conserved.

In our work, the aim of ontology mapping is different. We aim at creating a named entity resource, whose typology size is balanced with respect to NLP tasks such as named entity recognition (NER), machine translation, etc. This requires usually several dozens of types at most. Thus, we wish to map the types of

[7] `http://download.geonames.org/export/dump/allCountries.zip`

[8] `http://download.geonames.org/export/dump/PL.zip`

[9] `http://download.geonames.org/export/dump/alternateNames.zip`

our source ontologies (Wikipedia and GeoNames) on types and relations of Prolexbase so that only the typology of the latter resource is conserved. This mapping has been manually performed, as described in this section.

3.2.1 Mapping Wikipedia onto Prolexbase Ontology

All Polish Wikipedia classes (340 infobox classes or 676 person-related categories, cf Sec. 3.1.1) proved appropriate for a rather straightforward mapping onto appropriate Prolexbase types and existence values (historical, fictitious or religious). Lines 1–5 in Tab. 1 show some examples of such basic mappings. Moreover, numerous Wikipedia classes were specific enough to allow a global assignment of other relations as well. A (language-independent) meronymy relation with a toponym was the most frequent one. For instance (Tab. 1, line 8), the *Japońscy samuraje* 'Japanese samurai' class could be mapped on the *celebrity* type, *historical* existence, and *meronymy* relation with the holonym pivot representing *Japan*. Some categories allowed the assignment of more than one holonym (Tab. 1, line 11) or of an accessibility relation (Tab. 1, line 12).

The mapping and the selection of related pivots were done manually. As a result, each Wikipedia entry was automatically assigned the Prolexbase type, existence, meronymy and/or accessibility on which its Wikipedia class was mapped. Rare erroneous assignments that might result for individual entries from this global mapping were to be fixed in the human validation stage.

3.2.2 Mapping GeoNames onto Prolexbase Ontology

A mapping was also necessary between GeoNames and Prolexbase typologies. In most cases global assignment of GeoNames main categories to Prolexbase types was sufficient. However, several GeoNames subcategories refer to different Prolexbase types than their parent main categories, e.g. the subcategory *S.CAVE* (cave, caves) corresponds to the Prolexbase type *geonym* although its parent category *S* (spot, building, farm) is mapped on type *building*.

The two-level mapping of GeoNames categories and selected subcategories onto Prolexbase types is presented in Tab. 2 and Tab. 3.

3.3 Frequency Code Estimation

As mentioned in the Section 2.1, every language-specific entry (prolexeme) in Prolexbase obtains one of the three standard *frequency* values, which describes how popular the given prolexeme is:

1. commonly used,
2. infrequently used,
3. rarely used.

Since Wikipedia does not indicate any similar measure for its articles, we had to introduce our own routine for estimating this factor.

Initially, we examined two Wikipedia-inherent criteria: (i) the number of Wikipedia articles referring to the given article, (ii) the length of the article. It proved

Table 1: Sample Wikipedia classes mapped to Prolexbase types and relations

#	Wikipedia class name	Wikipedia class type	Prolexbase type	Prolexbase existence	Prolexbase relation	Prolexbase related entry	Prolexbase subject file
1	Jezioro 'Lake'	infobox	lake	historical			
2	Linia lotnicza 'Airline'	infobox	firm	historical			
3	Postać telenowela 'Soap opera character'	infobox	celebrity	fictitious			
4	Dziennikarze 'Journalists'	category	celebrity	historical			
5	Kompozytorzy XX wieku '20th century composers'	category	celebrity	historical			
6	MYS Stan 'State of Malaysia'	infobox	region	historical	meronymy	Malaysia	
7	Stacja polarna 'Research stations in Antarctica'	infobox	institution	historical	meronymy	Antarctica	
8	Japońscy samuraje 'Japanese samurai	category	celebrity	historical	meronymy	Japan	
9	Ludzie związani z Melbourne 'People from Melbourne'	category	celebrity	historical	meronymy	Melbourne	
10	Odznaczeni Orderem Łaźni 'Companions of the Order of the Bath'	category	celebrity	historical	meronymy	Order of the Bath	
11	Niemieccy teolodzy luterańscy 'German Lutheran theologians'	category	celebrity	historical	meronymy meronymy	Germany Lutheranism	
12	Władcy Blois 'Counts of Blois'	category	celebrity	historical	accessibility	Blois	leader

Table 2: GeoNames main categories mapped to Prolexbase types.

Symbol	Geonames Category	Prolexbase type	Symbol	Geonames Category	Prolexbase type
A	country, state, region	region	S	spot, building, farm	building
H	stream, lake, …	hydronym	T	mountain, hill, rock	geonym
L	parks, area, …	geonym	U	undersea	hydronym
P	city, village, …	city	V	forest, heath, …	geonym
R	road, railroad	way			

Table 3: GeoNames subcategories mapped to different Prolexbase types than their parent categories.

Symbol	Geonames Subcategory	Prolexbase type	Symbol	Geonames Subcategory	Prolexbase type
A.PCLI	independ. polit. entity	country	L.PRK	park	building
L.AMUS	amusement park	building	L.RES	reserve	city
L.CMN	common	building	L.RESF	forest reserve	building
L.CTRB	business center	building	L.RESN	nature reserve	building
L.DEVH	housing development	building	L.RESV	reservation	building
L.MILB	military base	building	L.RESW	wildlife reserve	building
L.LCTY	locality	building	S.CAVE	cave(s)	geonym

that the correlation between these two properties and the popularity of an entry was not sufficiently systematic (e.g. some rather unknown entries have very rich and well referred articles due to a contribution of passionate users).

Having discarded these two criteria we turned towards measuring the popularity of Wikipedia entries by counting, through a significant period of time, how many times the corresponding article was entered by people browsing Wikipedia. The choice of the criterion was arbitrary but it seems to fit the intuition that the most popular are those entries which people access (search for, read or modify) most often. In order to count the *Wikipedia hits* for the previously extracted entries we used a service[10] (built upon page view statistics for Wikimedia projects[11]) which stores the number of hits of every Wikipedia article in any month (since December, 2007). In order to reduce the impact of short-term high popularity phenomena we collected 12-month statistics from January 1st, 2010 to December 31st, 2010. These statistics had to be mapped on the three-value scale of the Prolexbase

[10] `http://stats.grok.se/`

[11] `http://dumps.wikimedia.org/other/pagecounts-raw/`

frequency attribute.

As it turned out, the most frequently visited Wikipedia articles were those that regarded people and cities. Thus, if we simply set the threshold of the three frequency classes using the absolute number of hits, most entries of types different than celebrity and city would be given the frequency value 2 or 3 although many of them definitely seem well known. For instance, Shakespeare's *Romeo i Julia* 'Romeo and Juliet' seems more universally known than the French wrestler *André the Giant* but the number of Polish Wikipedia hits for the former was significantly lower than for the latter.

As a result of this observation we decided to split Wikipedia entries into 4 subclasses: cities (that made about a half of all entries that we had collected), people (celebrities – approx. 25% of all entries), works and other entries. Hit count thresholds of frequency groups were experimentally set for every subclass separately:

- for *celebrity* and *work* subclasses: 10% of entries with the highest number of visits received code 1 (*commonly used*), next 30% got code 2 (*infrequently used*) and the rest was assigned code 3 (*rarely used*),
- for *city* and *other* subclasses: the first 4% received code 1, next 16% – code 2, and the rest – code 3.

Table 4: Most frequently accessed entries in 2010 in English and French Wikipedia, and their hit counts.

	English Wikipedia		**French Wikipedia**	
City	New York City	7,706,928	Paris	2,496,853
	London	7,208,939	New York	1,025,958
	Paris	4,299,035	Londres 'London'	732,437
Celebrity	Justin Bieber	20,218,891	Justin Bieber	2,165,053
	Lady Gaga	18,772,163	Victor Hugo	1,614,872
	Eminem	14,255,827	Michael Jackson	1,560,958
Work	Glee 'TV series'	20,079,786	Desperate Housewives 'TV series'	1,764,269
	Avatar 'film'	13,208,450	Dr House 'TV series'	1,099,244
	Inception 'film'	10,923,537	Dexter 'TV series'	941,337
Other	United States	34,180,907	France	3,388,647
	United Kingdom	16,493,483	Haiti	1,695,685
	Google	16,153,818	Etats-Unis 'United States'	1,618,089

Tab. 4 shows the three most frequently accessed entries in each of the four subclasses in French and English Wikipedia, while Tab. 5 contains the three leading entries in Polish Wikipedia for each of the four subclasses and each of the three frequency values. Interestingly enough, while 5 of the most popular entries (*New York*, *London*, *Paris*, *Justin Bieber*, and *United States*) are shared between French and English Wikipedia, only one of them (*Dr House*) belongs the top three Polish and French entries simultaneously, and none of them is shared between Polish and English.

Table 5: Frequency codes based on statistics on Polish Wikipedia hits. Three leading entries are given for each class.

	Code 1		Code 2		Code 3	
City	Warszawa 'Warsaw'	1,114,854	Hawr 'Le Havre'	14,980	Gmina Aleksandrów Łódzki (a Polish commune)	3,996
	Kraków 'Cracow'	694,382	Nowe	14,949	Svidník (a Slovak commune)	3,995
	Wrocław	511,005	Annopol	14,928	Gmina Główczyce (commune)	3,993
Celebrity	Lech Kaczyński	1,746,776	André the Giant	34,975	Angie Stone	8,599
	Fryderyk Chopin	1,580,473	Sun Zi	34,960	Mary Elizabeth Winstead	8,598
	Jan Paweł II 'John Paul II'	1,307,229	Liam Hemsworth	34,940	Gerd Binnig	8,597
Work	Dr House	893,676	Pitch Black	11,999	Oczy węża 'Snake Eyes'	2,499
	Hannah Montana	562,030	Nowy Scooby Doo	11,978	Trollz	2,499
	The Vampire Diaries	550,790	Baranek Shaun 'Shaun the Sheep'	11,965	Oświadczyny po irlandzku 'Leap Year'	2,499
Other	Polska 'Poland'	3,472,268	Elba	24,995	White City Stadium	2,399
	Unia Europejska 'European Union'	1,272,789	Lexus	24,991	Bitwa morska pod Prevezą 'Battle of Preveza'	2,397
	Niemcy 'Germany'	1,188,483	Bazylika Św. Pawła za Murami 'Basilica of Saint Paul Outside the Walls'	24,988	Wzgórze Wiktorii 'Victoria Peak'	2,397

Each entry selected from GeoNames also had to be assigned one of the three frequency values. To this end, we introduced an intermediate geographical frequency (*geogFrequency*) code with values 1, 2 and 3, like the frequency code above. All Polish geographical entries stemming from GeoNames, Wikipedia and/or Prolexbase were temporarily merged, which resulted in a set of 60,348 entries. All names of countries, European and Polish administrative division units, as well as capitals of these entities, were put at the beginning of the list (in order to fullfil the criterion of systematic selection of some name categories mentioned in Section 3.1.2). Other entries, provided that they existed in Wikipedia, were ordered by the number of hits of the corresponding Wikipedia articles. Entries extracted from GeoNames and non-existent in Wikipedia were considered in the lexicographic order. The ordered list obtained in this way was divided into three sublists. The first 10,000 entries were assigned *geogFrequency* code 1, the 10,000 following entries – code 2, and the rest – code 3. The entries of the first sublist were then proposed for human validation. If they existed in Wikipedia, their frequency code was suggested, otherwise their *geogFrequency* code was used. The

latter concerned 5,386 names including 1,852 cities, 1,707 regions, 976 buildings, 642 geonyms, 143 hydronyms, 42 ways and 24 countries (in terms of Prolexbase types). The few country names selected from GeoNames that were not found in Wikipedia included mostly synonyms (in terms of Prolexbase relations), e.g. *Koreańska Republika Ludowo-Demokratyczna* 'Democratic People's Republic of Korea' was found in GeoNames while *Korea Północna* 'North Korea' appeared in Wikipedia.

3.4 Pivot Selection

Data extracted from Wikipedia represent concepts and relations some of which may already be present in Prolexbase. Thus, the main challenge is to preserve the **uniqueness of concepts**, i.e. to select the proper (language-independent) pivot if the current concept is already present in Prolexbase, and to create a new pivot otherwise. The fact of working on three languages simultaneously greatly increases the reliability of this process. Recall that Prolexbase originally contained mostly French data. If new Polish or English data were to be examined separately, few hints would be available as to the pre-existence of adequate pivots. For instance, if Prolexbase already contains the prolexeme *Aix-la-Chapelle* with pivot 45579, it is hard to guess that the incoming Polish prolexeme *Akwizgran* should be attached to the same pivot. If, however, all three equivalents — *Aachen* (EN), *Aix-la-Chapelle* (FR) and *Akwizgran* (PL) are extracted from Wikipedia their matching with pivot 45579 is straightforward.

While selecting the most probable pivot, ProlexFeeder assumes that: (i) the current content of Prolexbase has the validated status, (ii) data added automatically have the non-validated status, (iii) while validating an entry we rely only on the already validated data. Due to homonymy and variation, comparing the Wikipedia entry with a prolexeme is not enough. At least three other sources of evidence may be exploited. Firstly, some homonyms can be distinguished by their type, e.g. the Wikipedia entry *Aleksander Newski* as a work (film) should not be mapped on the pivot of type celebrity. Secondly, a Wikipedia entry may be equal to an alias rather than a prolexeme of an existing pivot. For instance, the main entry in Example (3) ('Great Britain'), is shorter than its alias ('United Kingdom of Great Britain and Northern Ireland') in Wikipedia, conversely to Prolexbase, where the most complete name is usually chosen as the prolexeme. Thirdly, a common URL is a strong evidence of concept similarity.

Consider Tab. 6 showing a sample set of Wikipedia data resulting (except the *pivot* attribute) from the preprocessing described in the preceding sections. Fig. 4 sketches the algorithm of pivot selection for a new data set e. Its aim is to find each pivot p existing in Prolexbase such that, for each language l (PL, EN or FR), the data linked with p (if any) are similar to e. The similarity between e and p grows with the decreasing value of the **distance** function, which compares the lexemes, aliases, URLs and types of e and p in the given language. If e and p have the same lexemes and share either the URL or the type then the distance is considered equal to zero (lines 15 and 18). An incoming lexeme should not only be compared with existing prolexemes but with their aliases as well (line 16). Note, however, that bidirectional matching of lexemes and aliases between Wikipedia and Prolexbase is

Function getPivotCandidates(e) return *pivotList*
Input e: structure as in Tab. 6 //incoming entry
Output *pivotList*: ordered list of (p, d) with $p, d \in \mathbb{N}$ //proposed pivots and their distances from e

1. **begin**
2. **for each** $l \in \{PL, EN, FR\}$ **do**
3. $pivots.l \leftarrow \langle\rangle$ //empty list
4. **for each** $p \in allPivots$ **do**
5. **for each** $l \in \{PL, EN, FR\}$ **do**
6. **if** $distance(e, p, l) < 10$ **then**
7. $insertSorted(p, pivots.l)$ //insert the new pivot in the sorted candidates list
//merge three sorted candidate lists into one
8. $pivotList \leftarrow mergeSorted(pivots.PL, pivots.EN, pivots.FR)$
9. **if** $pivotList = \langle\epsilon\rangle$ **then** //no similar pivot found
10. $pivotList \leftarrow \langle(getNewPivot(), 0)\rangle$ //create a new pivot
11. **return** *pivotList*
12. **end**

Function distance(e, p, l) return d
Input e: structure as in Fig. 6 //incoming entry
p: pivot
$l \in \{PL, EN, FR\}$ //language
Output $d \in \{0, 1, 2, 3, 10\}$ //distance between e and p

13. **begin**
14. $d = 10$
15. **if** $e.l.lex = p.l.lex$ **then** $d \leftarrow 0$ //same lexeme
16. **else if** $e.l.lex \in p.l.aliases$ **then** $d \leftarrow 1$ //lexeme same as an alias
17. **else if** $e.l.url = p.l.url$ **then** $d \leftarrow 2$ //matching Wiki URL
18. **if** $d \leq 1$ **and** $e.l.url \neq p.l.url$ **and** $e.type \neq p.type$ **then** $d \leftarrow 3$
19. **return** d
20. **end**

Figure 4: Algorithm for selecting candidate pivots for a new incoming entry.

not always a good strategy. For instance, the redirects in Example (5) are former names ('Napoleon Square', 'Warka Square') of a square ('Warsaw Uprising Square in Warsaw'). Recall that in Prolexbase such variants are not considered as aliases but refer to different pivots (linked by the diacronic synonymy relation). Note also that, as soon as e and p share the same URL, they are considered as close even if their lexemes and aliases or types differ (lines 17–18).

The *distance* function is used to compare an incoming Wikipedia entry e with each pivot existing in Prolexbase (lines 4–6). For each of the three languages we get a sorted list of pivots which are similar to e (line 7). The three resulting lists are then merged (line 8) by taking two factors into account: (i) the rank of a pivot in each of the three lists, (ii) its memberhood in the intersections of these lists. If no similar pivot was found in any language a new pivot is proposed (line 9–10).

Table 6: Sample preprocessed Wikipedia data. The attributes represent: Wikipedia lexemes (*PL.lex*, *EN.lex*, *FR.lex*), number of Wikipedia hits in 2010 (*PL.hits*, *EN.hits*, *FR.hits*), frequency (*PL.freq*, *EN.freq*, *FR.freq*), Wikipedia page URL (*PL.url*, *EN.url*, *FR.url*), Wikipedia redirects proposed as aliases (*PL.aliases*, *EN.aliases*, *FR.aliases*), predicted Polish inflected forms (*PL.infl*), predicted pivot, existence, meronymy-related pivot *meroPivot*, and Prolexbase type.

Attribute	**Value**	**Attribute**	**Value**	**Attribute**	**Value**
PL.lex	Rzym	EN.lex	Rome	FR.lex	Rome
PL.hits	315,996	EN.hits	3,160,315	FR.hits	450,547
PL.freq	1	EN.freq	1	FR.freq	1
PL.url	pl.wikipedia.org/wiki/Rzym	EN.url	en.wikipedia.org/wiki/Rome	FR.url	fr.wikipedia.org/wiki/Rome
PL.aliases	*Wieczne miasto*	FR.aliases	*Ville Éternelle*, *Ville éternelle*	EN.aliases	*Capital of Italy*, *Castel Fusano*, *Città Eterna*, …
PL. infl	*Rzymu:sg:gen:m3*, *Rzymowi:sg:dat:m3* *Rzym:sg:acc:m3*, …	type	city	existence	historical
		meroPivot	none	pivot	42787

The pivots returned by the algorithm in Fig. 4 are proposed to a human validator as possible insertion points for new Wikipedia data, as discussed in Section 4. When the correct pivot has been selected by the lexicographer ProlexFeeder considers different strategies of merging the new incoming data with the data attached to this selected pivot. For instance, an incoming lexeme may take place of a missing prolexeme or it can become an alias of an existing prolexeme. The values of frequency, URL, aliases, inflected forms, existence, holonym/meronym, and type predicted for the incoming entry (cf. Tab. 6) may be either complementary or inconsistent with those of the selected pivot. In the latter case, the Prolexbase data are considered as more reliable but the user is notified about the conflict.

As far the insertion of a GeoNames entry is concerned, the entry is first straightforwardly matched with the extracted Polish Wikipedia entries. If an identical entry is found then its attributes become those of the GeoNames entry (except, possibly, the frequency code, cf Section 3.3). Otherwise it is considered that the GeoNames entry has no corresponding Wikipedia entry and thus many attributes of its structure shown in Fig. 6 become empty. Note that this matching process is less reliable than matching Wikipedia entries with Prolexbase. This is because a significant amount of GeoNames entities does not have translations to other languages, e.g. *Zala*, a Hungarian first-order administrative division unit, is represented in GeoNames with its Hungarian name only. Although there exist articles describing the same concept in Polish and English Wikipedia (*Komitat Zala* and *Zala County*, respectively) they could not be mapped on *Zala* alone. As a result, both the Wikipedia and the GeoNames entries were suggested as new Prolexbase entries with two different pivots. This problem occurred most often for regions (European administrative division units) extracted from GeoNames, many

of which were cited in the holonym country's language only. During the human validation, proper Polish, English and French equivalents were to be found manually for such names, which made the whole procedure highly time-consuming. Therefore, those region names that were hard to identify manually were left for a further stage of the project.

4 Human Validation

The aim of Prolexbase is to offer high-quality lexico-semantic data that have been manually validated. Thus, the results of the automatic data integration presented in Section 3 do not enter Prolexbase directly but are fed to a graphical user interface offered by ProlexFeeder. There, the lexicographer views new entries proposed by the automatic selection and integration process, validates, completes and/or deletes them. She can also browse the current content of Prolexbase in order to search for possible skipped or mismatched pivots and prolexemes.

Most often, the incoming entries are new to Prolexbase but sometimes they match existing pivots. In this case the data coming from external sources complete those already present. For instance, Fig. 5 shows fragments of the validation interface for the data describing the Italian capital. Prolexemes in the three languages are proposed, together with their Wikipedia URLs (which are usually new to Prolexbase). The pivot selection procedure (cf Section 3.4) has found the proper existing pivot. The proposed alias *Wieczne miasto* 'eternal city', which is a stylistic synonym from the Prolexbase point of view, has been manually transformed into a new prolexeme and attached to a different pivot. Derivations (*rzymianin, rzymianka* 'Rome inhabitant', *rzymski*, 'Roman') can be added manually, and the proposed inflected forms of the Polish prolexeme can be corrected or validated. Inflection of aliases and derivatives is not done for the moment as it is supposed to be modelled later via external grammatical resources. Changes made to the current entry can finally be either cancelled or validated. The whole entry can also be deleted or put aside in order to be processed later.

5 Evaluation

In order to estimate both the quality of the data integration process and the usability of the human validation interface, samples of Wikipedia entries of three different types were selected: celebrity, work and city, containing 500 entries each. A lexicographer was to process these samples type by type in the GUI, collect the statistics about wrongly proposed pivots and count the time spent on each sample. Tab. 7 shows the results of this experiment. A true positive is a pivot that has existed in Prolexbase and is correctly suggested for an incoming entry. A true negative happens when there is no pivot in Prolexbase corresponding to the incoming entry and the creation of a new pivot is correctly suggested. A false positive is an existing pivot that does not correspond to the incoming entry but is suggested. Finally, a false negative is an existing pivot which corresponds to the entry but which fails to be suggested (i.e. the creation of a new pivot is suggested instead). Type city has the largest number of true positives since initially Prolexbase con-

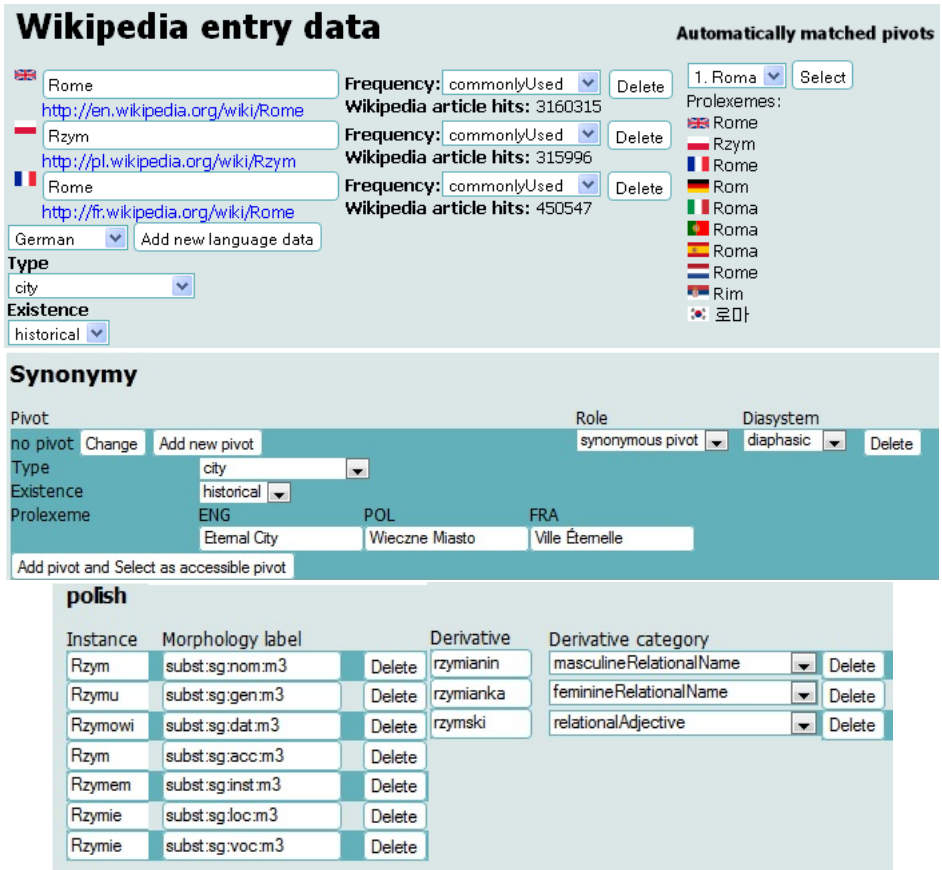

FIGURE 5: Fragments of the ProlexFeeder GUI for correcting and validating pivots, prolexemes, Wikipedia links, type, existence, frequency, synonyms, inflected forms and derivatives.

tained many French toponyms, some celebrity names and only very few names of works. The true negatives correspond to the effectively added new concepts. The false positives are infrequent and their detection is easy since the lexicographer directly views the details of the wrongly proposed pivot. False negatives are the most harmful since detecting them requires a manual browsing of Prolexbase in search of prolexemes similar to the current entry. Fortunately, these cases cover only 1.3% of all entries.

Wrongly selected pivots result mainly from the strict matching algorithm between an incoming lexeme and existing prolexemes and aliases (cf Fig. 4, lines 15–16). For instance, the Polish Wikipedia entry *Johann Sebastian Bach* did not match the Polish prolexeme *Jan Sebastian Bach*, while *The Rolling Stones* appeared in Prolexbase as *Rolling Stones* with a collocation link to *The*. Some true homonyms also appeared, e.g. the pivot proposed for *Muhammad Ali* as a boxer represented in fact the pasha of Egypt carrying the same name. The evidence of different French

Table 7: Results of ProlexFeeder on three sets of entries.

Type	Incoming entries	True positives	True negatives	False positives	False negatives	Accuracy	Workload
Celebrity	500	87	400	1	12	97.4%	21h30
Work	500	9	472	16	3	96,2%	17h30
City	500	226	264	6	4	98%	16h
All	1500	322	1136	23	19	97.2%	55h

equivalents (*Muhammad Ali* and *Méhémet-Ali*) was not sufficiently strong to allow the selection of different pivots. Similarly, *Leszno* in the Wielkopolska Province was mistaken for *Leszno* in Mazovia Province.

On average, the processing of an incoming entry takes about 2 minutes. Most of this time is taken by completing and/or correcting the inflected forms of Polish prolexemes (usually 7 forms for each name). Inflecting celebrity names proves the most labour-intensive since Translatica's automatic inflection tool (cf Section 2.4) makes some errors concerning person names: (i) their gender is wrongly guessed, (ii) the inflection of their components is unknown (thus we get e.g. **Maryla Rodowicza* instead of *Maryli Rodowicz*). Moreover the inflection of foreign family names is a challenge for Polish speakers.

The morphological description of works is easier since they often contain common words (*Skrzynia umarlaka* 'Dead Man's Chest') or they do not inflect at all (*Na Wspólnej* 'On the Wspolna Street'). The main challenge here is to determine the proper gender. For instance *Mistrz i Małgorzata* 'The Master and Margarita' may be used in feminine (while referring to the classifying context *książka* 'the book'), in masculine (the gender of *Mistrz* 'Master'), or even in masculine plural (to account for the coordination dominated by the masculine noun).

Inflecting city names proved relatively easy – most of them contained one word only and their morphology was rather obvious. Notable exceptions were again foreign names for which the application of a Polish inflection paradigm may be controversial (e.g. *w okolicach Viborga/Viborg* 'in the vicinity of Viborg'). Surprisingly enough, the major difficulty for this type came from the fact that almost 50% of the cities already had their pivot in Prolexbase. Since several settlements with the same name frequently occur checking all necessary relations in order to validate the suggested pivot could be non-trivial.

Other problems concerned types and relations. Wrong types were systematically proposed for some groups of Wikipedia entries due to particularities of Wikipedia categories and infobox types. For instance names of music bands (*Genesis*) are classified in Wikipedia jointly with individual celebrities, thus changing their Prolexbase type to Ensemble had to be done manually. In samples of type city only one type error appeared (*Trójmiasto* 'Tricity' had to be reclassified as a region), and all works had their type correctly set.

Missing relations are due to the fact that they are not directly deducible from the Wikipedia metadata that were taken into account up till now. Thus, the following relations had to be established manually: (i) accessibility between ensembles

and their members (*Wilki* and *Robert Gawliński*) or between works and their authors (*Tosca* and *Giacomo Puccini*), (ii) meronymy between celebrities or works and their birth or edition countries (*Kinga Rusin* and *Poland*, the *Wprost* magazine and *Poland*), (iii) meronymy between cities and countries or regions (if several settlements sharing the same name are situated in the same country the meronymy is established with respect to smaller territories allowing for semantic disambiguation). Recall also that derivatives had to be established fully manually.

Because Prolexbase models both semantic and morphological relations among proper names, we expect the benefit from this resource to be most visible in NLP applications dedicated to morphologically rich languages. The first estimation of this benefit has been performed for Nerf[12], a named entity recognition tool based on linear-chain conditional random fields. Nerf recognizes tree-like NE structures, i.e., containing recursively nested NEs. We used the named entity level of the manually annotated 1-million-word National Corpus of Polish[13] (Przepiórkowski *et al.*, 2012), divided randomly into a training and an evaluation part of 90% and 10% of words, respectively. Nerf has been trained once with no external resources (setting A), and once with the list of Polish Prolexbase instances and their types (setting B). Each setting admitted 20 training iterations. Evaluation has been performed on the level of separate tokens according to the IOB labeling principle (extended to nested NEs). The average accuracy of the models obtained after the 10 final iterations is equal to 0.97799 for setting A and to 0.97815 for setting B (note that the *O* labels attributed to tokens not belonging to named entities have an important contribution to this high score). The enhancement of performances due to the use of Polish Prolexbase data is slight but significant given the fact that the scores are close to 1.

6 Related Work

Prolexbase can be compared to **EuroWordNet** (Vossen, 1998) in that both organize multilingual lexica around a language-independent set of meaning identifiers, i.e. pivots in Prolexbase and Interlingual Index Records (ILIRs) in EuroWordNet. Both resources cross barriers between different parts of speech in that EuroWordNet connects e.g. verbs (*adorn*) with their derivatives (*adornment*), while Prolexbase attaches relative adjectives (*Roman*) to location names (*Rome*). There are, however, notable differences between both models. Firstly, Prolexbase is specialized in the representation of proper names, while EuroWordNet combines general-purpose wordnets. Secondly, each Prolexbase pivot is assigned a type, a supertype and an existence and is related with other pivots in the language-independent level. In EuroWordNet, conversely, ILIRs form an unstructured list of meanings whose sole purpose is mediation between synsets in language-specific wordnets. Only a very restricted set of ILIRs, representing the most general concepts, is assigned to the Top Ontology. Thirdly, the language-specific modules in Prolexbase, unlike in EuroWordNet, all follow the same hierarchical structure and express lexical rather than semantic relations. Finally, Prolexbase is an open

[12] http://zil.ipipan.waw.pl/Nerf

[13] `http://nkjp.pl/index.php?page=0&lang=1`

resource while EuroWordNet comes with a restrictive ELDA/ELRA license. Let's also note that EuroWordNet presently contains no Polish module.

Another related resource is the **Universal Wordnet** (UWN) (de Melo and Weikum, 2009), a graph-based extension of the Princeton Wordnet, thus containing general-purpose and partly domain-oriented vocabulary with no particular interest in named entities. It is available under the Creative Commons BY-NC-SA 3.0 license and contains 1,5 million meanings for 800,000 words from 200 languages. It results from an automatic compilation of various monolingual wordnets, translation dictionaries, multilingual and monolingual thesauri and ontologies, Wiktionary data and parallel corpora. The resulting unified graph is iteratively (2-4 times) enhanced in that new candidate links are infered from existing cross-lingual links, and the reliability of these new links is calculated by an SVM classifier trained on a limited set of randomly selected and manually validated links. The major principles of the semantic representation are similar to Prolexbase in that language-independent and language-specific layers are defined. The former consists in a set of uniform sense indentifiers mainly inspired by the Princeton WordNet. Sense frequency information is extracted from sense-annotated corpora. This can be seen as an alternative to Wikipedia hit counts in Prolexbase. It is unclear, however, if highly inflected, e.g. Slavic, languages could be concened by this measure, given the high variability of term surface forms in these languages. Concerning interlingual links the main difference is that Prolexbase provides translations via pivots, i.e. concepts, while UWN contains direct translation links between terms in different languages. We might think of UWN techniques as possible support for Prolexbase creation in that a big variety of versatile resources can be used on input. The main obstacle would probably come from the fact that, as disscussed later on, proper names are concepts in Prolexbase while they are instances in WordNet. Thus, adding a new (conceptual) proper name to Prolexbase implies enlarging its conceptual hierarchy, which does not seem to be possible with the UWN algorithms. This is crucial in extending Prolexbase both for languages already covered and for new language modules. Namely, each new language brings not only the translations for already existing concepts but also a whole range of new names (thus, concepts) used frequently in this language but rarely in others (e.g. names of nationally known people, places, institutions, etc.).

Prolexbase population from Wikipedia and GeoNames can be seen as an instance of the **ontology learning** problem, as discussed by Petasis *et al.* (2011). The notion of ontology is seen there rather broadly as "a formal explicit specification of a shared conceptualization" (as opposed to a **taxonomy**, in which concepts are necessarily organized hierarchically by the subsumption, i.e. "is-a", relation). Our approach has, however, an essential specificity which comes from the status proper names are given in Prolexbase. Recall that proper names are represented by **pivots** in the language-independent layer and by **prolexemes** and their variants in each language module. In other words, proper names correspond both to **concepts**, called **conceptual proper names** by Prolexbase authors (Krstev *et al.*, 2005) and to **instances** of the Prolexbase ontology. Thus, we simultaneously perform **ontology enrichment** (placing new concepts and relations at the correct position in an existing ontology) and **ontology population** (adding new instances of existing concepts). According to the taxonomy of ontology learning methods proposed

by Petasis *et al.* (2011), our ontology enrichment is based on **integrating** existing ontologies (as opposed to constructing an ontology from scratch and specializing a generic ontology). Our ontology population, in its turn, is atypical since we use instances of existing ontologies and inflection tools, rather than extraction from text corpora.

Ontology integration corresponds roughly to what Shvaiko and Euzenat (2013) call **ontology matching**. Our position with respect to the state of the art in this domain is twofold.

Firstly, we perform a mapping of Wikipedia classes and GeoNames categories on Prolexbase types (cf. Section 3.2). This process is not central to our work, and we have not designed it so as to optimize its generalization to other resources. With respect to the taxonomy of the methods reviewed in (Shvaiko and Euzenat, 2013), this mapping process can be described as producing **equivalence relations** (as opposed to subsumption relations) and resulting in a particular type of an **1:n alignment**. Namely, a Wikipedia infobox class is mapped on one Prolexbase type and on a set of relations (cf. Tab. 1). Our alignment is also fully manual, as opposed to all other methods cited in Shvaiko and Euzenat (2013), in which automating the ontology matching process is the main objective. Note also that instance-based ontology matching approaches, mentioned in the same survey, can be seen as opposed to ours. They use instances attached to concepts as evidence of concept equivalence, while we, conversely, rely on the types of proper names (i.e. concepts) from Wikipedia or GeoNames in order to find the equivalent names (i.e. instances), if any, in Prolexbase.

Secondly, we map names from Wikipedia and GeoNames on **conceptual proper names** (pivots) in Prolexbase (cf. Section 3.4). According to the taxonomy in (Shvaiko and Euzenat, 2013), this mapping is inherently multilingual (which is rare in other ontology matching methods) and is based on equivalence relation. It outputs 1:1 alignments, has a full-fledged matching validation interface (infrequent in other methods), and it performs the operation of ontology merging (as opposed to question answering). It uses string equality on the terminological level, is-a-similarity on the structural level, object similarity on the extensional level and does not apply any method on the semantic level.

This comparison with ontology matching state of the art is not quite straightforward since no conceptualization of proper names takes place in Wikipedia and GeoNames (but also in other common ontologies, like WordNet). Thus, mapping **multilingual sets of instances** (names) from Wikipedia and GeoNames on Prolexbase **pivots** corresponds to an **instance-to-concept** rather than a **concept-to-concept** matching.

This is why our method can more easily be situated with respect to the problem of the **creation and enrichment of lexical and semantic resources**, in particular for proper names, and their alignment with free encyclopaedia and thesauri. This problem has a rather rich bibliography most of which has been dedicated to English.

Toral *et al.* (2008) aim at an automatic extension of Princeton **WordNet** with named entities (NEs). A two-stage process is applied:

1. Synset-to-category mapping — each noun WordNet synset *s* containing classes

(not instances) is mapped on a set of one or more Wikipedia categories *C* by lemma matching. In this process polysemous words contained in *s* are disambiguated in that their instances are matched against Wikipedia articles belonging to categories in *C*.

2. NE selection — the entries belonging to the categories in *C*, and their hyponymy-related subcategories, are extracted. Capitalization norms in 9 languages are used to determine if a selected entry is a named entity or not. Namely, it is considered NE if it is capitalized in at least 91% of occurrences in the corresponding articles' bodies (in 9 languages). The NEs retained in this way are incorporated into the English WordNet as new synsets linked to *s* via the instance-of relation.

Some problems related to the synset-to-category mapping are mentioned: (i) a synset corresponds to a Wikipedia article rather than to a Wikipedia category, (ii) there is no category or article representing the same concept as the synset, (iii) the lemma-based mapping cannot be performed due to a tagger error or term variants, (iv) the synset nouns are polysemous and disambiguation fails due to non-existence of common instances.

In (Toral *et al.*, 2012) the authors propose refinements of their initial method and extend it to three languages: English, Spanish and Italian. Disambiguation of polysemous words is enhanced with: (i) extension of the instance mapping to hyponymy-related Wikipedia subcategories, (ii) text similarity measures between WordNet glosses and abstracts of the mapped Wikipedia categories. Identification of named entities in the selected categories is refined in that for a NE candidate entry *e* salient terms are extracted from the Wikipedia article describing *e*. These salient terms, together with *e*, become queries to a Web search engine; the top-ranked pages are used to search for capitalised vs. non-capitalised spelling of *e* in nine languages. If the proportion exceeds a threshold, the entry is considered a NE. The resulting tri-lingual NE lexicon contains almost a million English, over 137,000 Spanish and over 125,000 Italian fully automatically extracted NEs. Subsets of those NEs are linked to the English WordNet, SUMO ontology, and SIMPLE Italian computational lexicon. The lexicon is exportable in an LMF-compatible standard format.

The method is claimed to be applicable to any language provided that Wikipedia, a language resource with a noun taxonomy and a lemmatizer, exist for this language. We think nevertheless that the usefulness of this method is rather limited for the construction of a highly reliable resource such as the one that we are trying to obtain. Namely, it cannot be used as an iterative enrichment process. It seems that the authors never consider the problem of possible duplication of concepts or instances induced by an automatic extraction of entries and linking them to existing resources (recall that this problem is central to our considerations, cf. Sec. 3.4). In other words, they extract all entries (considered NEs) contained in the selected categories and do not compare them to those NEs that may have already existed in the resources to be enriched. This is probably why the newly extracted entries are never really integrated into those resources but they form instead a separate NE lexicon partly linked to those initial resources. An important quality problem appears here from our point of view. Different extraction

stages show accuracy below 78%, precision below 80% or recall below 91%. Even if these results can be considered good for a fully automatic process, a human validation and correction of results are needed if a high-quality subset of the resource is required. One cannot, however, benefit from such a costly validation in the following update of the NE lexicon since new results cannot be merged with the previous ones. The second reason of the limited usefulness of this method for our model concerns relations. Toral *et al.* (2012) only extract instance-of relations, while Prolexbase contains a richer, proper name-targeted and NLP-oriented set of relations.

Fernando and Stevenson (2012) deal with automatic mapping of English **WordNet** noun synsets onto Wikipedia articles. The whole contents of English Wikipedia (including the articles' contents) is indexed within an Information Retrieval (IR) system. For a given noun synset *S* the mapping procedure consists of three steps:

1. Candidate selection — it exploits the whole content of Wikipedia in order to reduce the search space from 3 million to less than 20 articles, while preserving 96% recall:
   - Words in *S* are matched against titles and redirects in Wikipedia and the corresponding articles are retained.
   - Lemmas of the words in *S* and in its gloss are used as queries to the IR system. Ten top ranked Wikipedia articles are retained.
2. Best candidate selection — the candidate articles retained in the previous step are normalized (by stemming, removing markup and stopwords). Text similarity between lemmas of the words in *S* and each normalized article is calculated (an article's title is given the highest weight in the similarity measure). The most similar article is retained only if its similarity score exceeds an experimentally determined threshold. Otherwise it is considered that there is no article representing the same concept as the synset.
3. Mapping refinement — mappings where more than one synset is aligned with the same Wikipedia article are removed; inter-article links are exploited to remove other unreliable mappings.

The resulting alignment allows retrieving new untyped relations between synsets based on Wikipedia links. These untyped relations are then used to enhance graph-based word sense disambiguation. Experiments with a set of 200 manually mapped noun synsets show that: (i) for 63% of synsets there exists a Wikipedia article describing the same concept, (ii) the mapping method reaches a precision of 87.8% and a recall of 46.9%, (iii) additional techniques (use of lemmas and glosses in synsets related to *S*, and considering articles referred to by disambiguation links) decrease the mapping quality.

Note that in this approach mapping is performed from synsets to Wikipedia articles, i.e. in the opposite direction than in our approach (from Wikipedia articles to pivots). Thus, it could inspire a new Prolexbase enrichment process in which the current set of pivots is considered as relatively complete (for a certain application) and data for a new language are to be added from Wikipedia.

A mapping in the same direction, i.e. from concepts to Wikipedia articles, is performed for English by Nguyen and Cao (2010). The longest name contained in an ontology entity (possibly reduced by elimination of suffixes such as *inc.*, *Mr.*, or *company*, etc.) is used as a Wikipedia query. The retrieved articles are ranked by a similarity measure based on bag-of-word feature vectors consisting of: Wikipedia article titles, redirects, categories, as well as incoming and outcoming links. The enriched proper name **KIM** ontology is used for iterative and incremental proper name disambiguation.

Freebase (Bollacker *et al.*, 2007) is as graph-shaped database of structured general human knowledge. It is inspired by Semantic Web and Wikipedia, i.e. it tries to merge the scalability of structured databases with the diversity of collaborative wikis. Rather than a system of rigid ontologies, it uses collaborative design of simple properties and types (with no type hierarchy). Conflicting and contradictory types and properties may exist simultaneously in order to reflect users' differing opinions and understanding. Freebase contents is provided by initial seeding of general-purpose data sets (culture, locations, science and Wikipedia knowledge). It is further extended by human collaborative efforts supported by a versatile Web interface allowing for: (i) automatic graph-based structural data matching, (ii) approximate string matching of literals, (iii) "undo" operations due to versioning. Freebase contains currently[14] 38,206,366 topics (people, works abstract concepts, etc.) and 1,180,485,054 facts. All the data are expressed in English and available under the Creative Commons Attribution CC-BY license.

YAGO (Suchanek *et al.*, 2007) is an ontology extracted automatically from the English Wikipedia and unified with WordNet. It has a well-defined logic-based knowledge representation model in which entities, facts, relations between facts and properties of relations are expressed in the *Yago model*, which is a slight extension of RDFS, a language for representing simple RDF (*subject-predicate-object* triple) vocabularies on the Web. YAGO population is based on Wikipedia *leaf* categories for entities (e.g. *Albert Einstein*). *Conceptual* leaf categories (e.g. *Naturalized citizens of the United States*) are automatically distinguished from *thematic* categories (e.g. *Physics*) via shallow parsing (a category is considered conceptual if the head noun is in plural, e.g. *citizens*). Head compounds (head nouns with possible modifiers) are then straightforwardly mapped on WordNet synsets (the most frequent synset is picked in case of ambiguities). Relational and other Wikipedia categories (e.g. *1879_births*) yield facts (e.g. *bornInYear*).

In (Hoffart *et al.*, 2011) YAGO2 extends YAGO by particularly focusing on temporal and spatial knowledge: (i) existence time spans are added to 76% of all entities, (ii) geographical coordinates and meronymy relations, extracted from Wikipedia and GeoNames, are assigned to all location entities. In this process Wikipedia and GeoNames entities are mapped on each other through both textual similarity and geographic proximity. GeoNames classes, in their turn, are mapped on YAGO by slightly enhanced original shallow noun phrase parsing. Multilingual translations are accessible via interwiki links for entity names, and via links to *Universal Wordnet* for classes. RDF triples expressing facts are extended into 6-tuples (*subject-predicate-object-time-location-context*) called SPOTLX. Spacio-

[14] According to `http://www.freebase.com/` accessed on March 24, 2013.

temporal questions (e.g. *companies founded nearby San Francisco*) can be expressed in a dialect of SPARQL, an SQL-like query language for RDF.

MENTA (de Melo and Weikum, 2010) is a named entity taxonomy compiled automatically from editions of Wikipedia in 200 languages. The construction algorithm is based on heuristic linking functions that connect Wikipedia articles, categories, infoboxes and WordNet synsets from multiple languages. These linkings are further refined by: (i) clustering multilingual entity names, (ii) ranking subsumption relations. The resulting taxonomy tends to account for culture-specific entities and categories.

DBpedia[15] (Bizer *et al.*, 2009; Mendes *et al.*, 2012) is dedicated to building Semantic Web by extracting structured information from Wikipedia in 97 languages. Data are extracted from an article's infobox, geo-coordinates, links to external sources, disambiguation pages, redirects and interwiki links. An atomic piece of extracted data is represented as an RDF triple and data queries can be expressed in SPARQL. Data are updated both after new Wikipedia dump releases and via real-time Wikipedia update report service. One of DBpedia's most valuable outcomes is a mapping of Wikipedia infobox templates and their properties on an interlingual **DBpedia Ontology** consisting of 359 classes and 1,775 properties[16]. This allows to alleviate inconsistent use of Wikipedia infoboxes. The mapping is being performed manually by a collaborative effort of volunteers in 24 languages. English and Polish Wikipedias belong to the 5 language versions whose mapping on the DBpedia ontology exceeds 80% of template occurrences. For French, 42% template occurrences are mapped. Note that such infobox template mapping might be supported by fully automatic methods (Nguyen *et al.*, 2011) in which similarity of infobox attributes can rely on their correlation and on vector similarity of their possible values. DBpedia RDFs contain outgoing links to other data sources, notably GeoNames and YAGO, and, conversely, increasingly many data publishers set RDF links to DBpedia, which makes DBpedia one of the central interlinking hubs in the Web of Data (network of interlinked data sources).

DBpedia also provides data sets explicitly created to support natural language processing tasks. The Lexicalization Data Set, built from Wikipedia titles, redirects, disambiguation links and anchor texts, contains alternative surface forms for a name and their association strength with a given concept or entity (identified by a URI). The Topics Signatures Data Set links each entity with a set of its most relevant terms extracted from article paragraphs containing a link to this entity. The Thematic Concepts Data Set links Wikipedia categories with DBpedia entities/concepts which represent the main themes of these categories. The Grammatical Gender Data Set, based on naive extraction of gender-specific pronouns, marks each entity mapped to the Person class as male or female.

Tab. 8 shows a contrastive study of methods dedicated to extraction and enrichment of structured data from Wikipedia. As can be seen, we offer one of the approaches which explicitly focus on modelling proper names instead of all nominal or other entities and concepts. Like YAGO and Freebase authors, but unlike others, we use multiple knowledge sources, and like two other approaches we con-

[15] `http://dbpedia.org`
[16] On November 16, 2012

sider several languages simultaneously rather than English alone. We share with DBpedia the idea of a manual typology mapping from Wikipedia infobox templates to ontology types, but we extend the relative (with respect to categories) reliability of infobox assignment by including articles from categories automatically judged as reliable. Like Freebase[17] but unlike all others we manually validate all preprocessed data. Most important, we aim at a limited size but high quality, manually validated, resource explicitly dedicated to natural language processing and focused on proper names. Thus, we are the only ones to:

- consider proper names as concepts of our ontology, which results in non-standard instance-to-concept matching,
- select and order input entries by popularity (estimated via Wikipedia hits),
- describe the full inflection paradigms for the retrieved names (notably for Polish being a highly inflected language),
- associate names not only with their variants but with derivations as well.

We also inherit Prolexbase's novel idea of synonymy in which a (diachronic, diaphasic or diastratic) change in the point of view on an entity yields a different although synonymous entity (note that e.g. in WordNet synonyms belong to the same synset and thus refer to the same entity). This fact enables a higher quality of proper name translation in that a synonym of a certain type is straightforwardly linked to its equivalent of the same type in another language.

Last but not least, ProlexFeeder seems to be the only approach in which the problem of a proper integration of previously existing and newly extracted data (notably by avoiding duplicates) is explicitly addressed. Thus, we truly propose an enrichment of a pre-existing proper name model rather than its extraction from scratch.

Wikipedia is the main sources of data for ontology creation and enrichment in the methods discussed above. An opposed point of view is represented by the **Knowledge Base Population**[18] (KBP) task defined within the Text Analysis Conference[19] (TAC), successor of TREC[20], organized since 2008 and dedicated to providing the infrastructure for large-scale evaluation of NLP technology. The KBP task aims at boosting technologies for building and populating knowledge bases (KBs) about named entities from unstructured text (mainly newswire). In particular, its 2011 mono-lingual (English) and cross-lingual (Chinese-English[21]) **Entity Linking** track is partly relevant to our work. In this track, the initial KB consists of about 818,000 entities extracted from English Wikipedia. Each entity is annotated with one of four types (person, organization, geo-political entity or unknown), with several (possibly empty) slots roughly equivalent to Wikipedia infobox attributes, and with a disambiguating text stemming from the Wikipedia article. Given a named entity and a source text (in English for the mono-lingual

[17] We have not found any information about the proportion of truly manually validated Freebase data (as opposed to the initial seeding data, whose validation method is unclear).

[18] `http://www.nist.gov/tac/2013/KBP/index.html`

[19] `http://www.nist.gov/tac/about/index.html`

[20] `http://trec.nist.gov/`

[21] In 2012 the set of languages has been extended with Spanish. Proceeding and results from this edition are not yet available.

TABLE 8: Contrastive analysis of approaches dedicated to extraction and enrichment of structured data from Wikipedia

Reference	Scope	Data Sources	Target Resource	Ontology Mapping Method	Ontology Mapping Source Unit	Ontology Mapping Target Unit	Population Method	Popularity Estimation Source	New Entry's Linguistic Features	Entry Validation Method	Languages	# entries
(Bollacker *et al.*, 2007)	unrestricted	Wikipedia, user's knowledge	Freebase	semi-manual collaborative, concept-to-concept	versatile	Freebase non-hierarchical types	semi-manual collaborative	none	none	manual, collaborative	English	38M entities 1,180M facts
(Suchanek *et al.*, 2007) (Hoffart *et al.*, 2011)	unrestricted	Wikipedia, WordNet, GeoNames	YAGO	automatic, concept-to-concept	Wikipedia leaf category, GeoNames classes	WordNet synset	extracting entries & facts	none	synonyms, variants	none	English	10M entities 120M facts
(Bizer *et al.*, 2009) (Mendes *at al.*, 2012)	unrestricted	Wikipedia	DBpedia	manual, collaborative & automatic, concept-to-concept	Wikipedia infobox template & attribute	DBpedia Ontology class & property	extracting & updating entries	none	variants, terms, themes & grammatical gender	none	24 languages	5.4M
(Nguyen & Cao, 2010)	proper names	Wikipedia	KIM	automatic, concept-to-concept	WordNet synset	Wikipedia article	adding features	none	NA	NA	English	unknown
(Melo & Weikum, 2010)	proper names	Wikipedia WordNets	MENTA	automatic ontology induction			extracting entries & relations from scratch	none	synonyms, variants	none	200 languages	unknown
(Toral *et al.*, 2012)	proper names	Wikipedia	LMF lexicon	automatic concept-to-concept	WordNet synset	Wikipedia category	extracting entries & relations from scratch	none	links with SIMPLE lexicon for Italian	automatic: capitalization rules, salient terms, Web search	English, Spanish, Italian	1M EN, 137K SP, 125K IT
(Fernando & Stevenson, 2012)	nouns	Wikipedia	WordNet	automatic concept-to-instance	WordNet synset	Wikipedia article	adding untyped relations	none	NA	NA	English	156K relations
Our approach	proper names	Wikipedia, GeoNames, Translatica	Prolexbase	manual concept-to-concept, semi-manual instance-to-concept	Wikipedia infobox template, GeoNames category, instances	Prolexbase type, relation and pivot	adding entries, relations & features	Wikipedia hits	inflection, variation, derivation	manual	Polish, English, French	39K PL, 33K EN, 100K FR

track, and in English or Chinese for the cross-lingual one) in which it appears, the task is to provide the identifier of the same entity in the KB, or NIL if this entity does not appear in the KB. All non-KB (NIL) entities have to be clustered in order to allow for the KB population. This task is similar to the pivot selection process in ProlexFeeder except that the typology is very light, the source languages are not concerned with high morphological variability in texts and, most important, entity mapping evidence is found in a corpus rather than in an existing, already structured, ontology. Sample TAC KBP results of the 2011 cross-language entity linking evaluation spread from 0.386 to 0.809 in terms of the B-cubed F-score.

Another TAC KBP track is **Slot Filling**. Given an entity name (person or organization), its type, a document in which it appears, its identifier in the KB, and a certain number of slots, the task is to fill these slots with data extracted from the document. This partly resembles the process of populating relations in ProlexFeeder. However, unlike relations in Prolexbase, the KBP track slots are flat labels or values rather than virtual relations to other existing KB nodes.

The above state of the art mentions only some major initiatives in creation and enrichment of lexical and semantic resources. Many other efforts have been made towards the construction of particular application- or language-oriented proper name thesauri and their exhaustive study is out of the scope of our paper. **JRC-NAMES** (Steinberger *et al.*, 2011) is a notable example in which a lightly structured thesaurus of several hundreds of thousands of named entities, mainly person names, is being continuously developed for 20 languages. New names are extracted by a rule-based named-entity recognizer from 100,000 news articles per day. Some manual validation is performed for the most frequently occuring names. Their spelling and transliteration variants are automatically detected by similarity measures, as well as by human-controlled Wikipedia mining. As a result, light is shed on the high variability of person names (30% of them have at least two variants, almost 2% have 10 or more variants, and 37 names have more than 100 variants). Morphological variants are also partly accounted for in that they are generated by hand-crafted rules and then matched against corpus occurrences.

Conclusions and Perspectives

We have described resources, methods and tools used for an automated enrichment of Prolexbase, a fine-grained high-quality multilingual lexical semantic resource of proper names. Three languages, Polish, English and French, were studied. The initial data contained mainly French names. New data were extracted from Wikipedia and GeoNames, and their integration with Prolexbase was based on a manual mapping of the three corresponding typologies. Attention was paid to establishing the degree of popularity of names, represented by their automatically pre-calculated frequency value, based in particular on Wikipedia hits of the corresponding entries. The morphological description of Polish names was supported by automatic inflection tools. The results of these preprocessing tasks were fed to ProlexFeeder, which contains two main modules: the pivot mapping, which automatically finds the proper insertion point for a new entry, and the graphical

lexicographer's interface, which enables a manual correction and validation of data.

Two main challenges in this automated data integration process are: (i) preserving the uniqueness of concepts, which are represented in Prolexbase by pivots, i.e. pairs of objects and points of view on these objects, (ii) offering a user-friendly and efficient lexicographer's workbench. Our experimental study has shown that over 97% of pivots proposed automatically by ProlexFeeder for the new incoming data are correctly identified. The lexicographer needs about 2 minutes to process an entry in the validation interface. The most challenging subtask is the Polish inflection of foreign names.

Tab. 9 shows the state of Prolexbase at the end of October 2012. The dominating role of toponyms is due to the initial contents of Prolexbase, which essentially focused on French geographical names. The most numerous types are city (46,677 pivots), celebrity (7,978 pivots), hydronym (4,554 pivots) and region (3,177 pivots), the number of pivots of the remaining types is between 1 and 970. Recall that one of original aspects of Prolexbase is the synonymy relation between pivots referring to the same object from different points of view. Currently, 3.35% of all pivots, mainly celebrities and countries, are in synonymy relation to other pivots. Moreover, about 89% and 8% of pivots are concerned by meronymy and accessibility relations, respectively.

Table 9: Current state of Prolexbase. Polish instances include inflected forms of prolexemes only.

Pivots				
All	Toponyms	Anthroponyms	Ergonyms	Pragmonyms
67,074	84%	14%	1,5%	0,5%

Relations			
All	Meronymy	Accessibility	Synonymy
65,494	92%	6%	2%

	Pivots in synonymy relation	**Pivots in meronymy relation**	**Pivots in acessibility relation**
All	2,453 (3.35%)	65,655 (89.59%)	6,312 (8.61%)
Most frequent types	celebrity 1,325 (16.61%) country 390 (44.62%) city 157 (0.32%)	city 48,104 (99.52%) celebrity 7,053 (88.39%) region 4,026 (96.71%)	city 2,214 (4.58%) region 1,696 (40.74%) celebrity 1,129 (14.15%)

Language	**Prolexemes**	**Aliases**	**Derivatives**	**Instances**
PL	27,274	8,664	3,083	165,324
EN	19,357	13,906	94	18,443
FR	70,764	8,440	20,919	142,393

All Prolexbase data are available from [anonymous link] under the CC BY-SA

license[22], i.e. the same as for Wikipedia and GeoNames. We are currently working on their LMF exchange format according to Bouchou and Maurel (2008).

Prolexbase is an open-ended project. Currently we have almost finished the processing of the names estimated as commonly used. This estimation was based on Wikipedia frequency data for 2010, and on GeoNames classification. Since both the contents of these two resources and the popularity of some names (e.g. those of celebrities and works, as seen in Tab. 4) are evolving, the Prolexbase frequency values deserve updates, possibly based on larger time intervals. Moreover, now, that the morphosyntactic variability of many names (in particular in Polish) has been described via instances, additional evidence of a name's popularity might stem from it's corpus frequency.

Note also that only a part of the relations modelled in Prolexbase has been actually dealt with in ProlexFeeder. The remaining linguistic-level relations, such as sorting order, collocations, classifying contexts, etc. are still to be described. Pragmonyms and ergonyms are underrepresented and should be completed. Instances are awaiting an intentional description, possibly encompassing both inflection and word formation (creating aliases and derivatives from prolexemes) within the same framework. It should, in an ideal case, be integrated with open state-of-the-art Polish inflection resources such as *PoliMorf*[23].

In order to ensure an even better pivot selection process, matching prolexemes and aliases could be enhanced by approximate string matching methods. Moreover the preprocessing methods might extend the scope of the automatically predicted relations by integrating approaches which exploit the internal structure of infoboxes and mine free text contained in Wikipedia pages.

We also plan to develop a more powerful Prolexbase browser within the ProlexFeeder's user interface. Multi-criteria search, as well as efficient visualisation and navigation facilities would greatly enhance the usability of the tool.

New development is also planned for the Prolexbase model itself. Firstly, a better representation of metonymy is needed. Recall (Section 2.1) that systematic metonymy (e.g. the fact that any city can be seen as a toponym, and anthroponym or a pragmonym) is currently expressed at the conceptual level by the secondary typology. However, some types are concerned by metonymy on a large but not systematic basis. For instance many names of buildings can refer to institutions they contain (*Muzeum Narodowe* 'The National Museum') but it is not always the case since a building can contain several institutions (*Pałac Kultury* 'The Palace of Culture').

Important challenges also concern the representation of the internal structure of multi-word proper names, seen as particular cases of multi-word expressions (MWEs). Recent development in applications such as coreference resolution, corpus annotation and parsing show that enhancement in lexicon/grammar interface is needed with respect to MWEs. For instance, the multi-level annotated National Corpus of Polish represents both named entities and syntactic groups as trees (Author et al. 2010), (Przepiórkowski *et al.*, 2012). Human or automatic annotation of such a corpus can greatly benefit from a rich linguistic resource

[22] http://creativecommons.org/licenses/by-sa/3.0/

[23] `http://zil.ipipan.waw.pl/PoliMorf`

of proper names such as Prolexbase. However, multi-word names contained in such as resource should possibly already be described as trees that could be reproduced over the relevant occurrences in the corpus. At least two kinds of trees are needed: (i) syntactic parse trees, (ii) semantic trees whose nodes are names embedded in the given name (e.g. *[[Wydział Teologii]$_{orgName}$ [Instytutu Katolickiego w [Paryżu]$_{settlement}$]$_{orgName}$]$_{orgName}$* '[[Faculty of Theology]$_{orgName}$ of the [Catholic Institute in [Paris]$_{settlement}$]$_{orgName}$]$_{orgName}$'). An efficient representation of such trees within Prolexbase is one of our major perspectives.

Finally, linking Prolexbase to other knowledge bases such as DBpedia or YAGO would combine the Semantic Web modelling benefits with advanced natural-language processing-oriented features and allow interlinking Prolexbase with many other data sets.

References

Claire Agafonov, Thierry Grass, Denis Maurel, Nathalie Rossi-Gensane, and Agata Savary (2006), La traduction multilingue des noms propres dans PROLEX, *Meta*, 51(4):622–636, les Presses de l'Université de Montréal.

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann (2009), DBpedia - A crystallization point for the Web of Data, *J. Web Sem.*, 7(3):154–165.

Kurt Bollacker, Patrick Tufts, Tomi Pierce, and Robert Cook (2007), A Platform for Scalable, Collaborative, Structured Information Integration, in *Proceeding of the Sixth International Workshop on Information Integration on the Web.*

Béatrice Bouchou and Denis Maurel (2008), Prolexbase et LMF : vers un standard pour les ressources lexicales sur les noms propres, *TAL*, 49(1):61–88.

Gerard de Melo and Gerhard Weikum (2009), Towards a universal wordnet by learning from combined evidence, in *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, November 2-6, 2009*, pp. 513–522, ACM.

Gerard de Melo and Gerhard Weikum (2010), MENTA: inducing multilingual taxonomies from wikipedia, in *Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, October 26-30, 2010*, pp. 1099–1108, ACM.

Samuel Fernando and Mark Stevenson (2012), Mapping WordNet synsets to Wikipedia articles, in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Sergio Ferrández, Antonio Toral, Óscar Ferrández, Antonio Ferrández, and Rafael Muñoz (2007), Applying Wikipedia's Multilingual Knowledge to Cross-Lingual Question Answering, in *Proceedings of the 12th International Conference on Applications of Natural Language to Information Systems, NLDB 2007*, volume 4592 of *Lecture Notes in Computer Science*, pp. 352–363, Springer.

Filip Graliński, Krzysztof Jassem, and Michał Marcińczuk (2009), An Environment for Named Entity Recognition and Translation, in *Proceedings of the 13th Annual Meeting of the European Association for Machine Translation (EAMT'09)*, pp. 88–96, Barcelona.

Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, Edwin Lewis-Kelham, Gerard de Melo, and Gerhard Weikum (2011), YAGO2: exploring and querying world knowledge in time, space, context, and many languages, in *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011 (Companion Volume)*, pp. 229–232, ACM.

Krzysztof Jassem (2004), Applying Oxford-PWN English-Polish dictionary to Machine Translation, in *Proceedings of 9th European Association for Machine Translation Workshop, "Broadening horizons of machine translation and its applications", Malta, April*, pp. 98–105.

Valentin Jijkoun, Mahboob Alam Khalid, Maarten Marx, and Maarten de Rijke (2008), Named entity normalization in user generated content, in *Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data, AND 2008*, ACM International Conference Proceeding Series, pp. 23–30, ACM.

Cvetana Krstev, Duško Vitas, Denis Maurel, and Mickaël Tran (2005), Multilingual Ontology of Proper Names, in *Proceedings of Language and Technology Conference (LTC'05), Poznań, Poland*, pp. 116–119, Wydawnictwo Poznańskie.

Giridhar Kumaran and James Allan (2004), Text classification and named entities for new event detection, in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pp. 297–304.

Denis Maurel (2008), Prolexbase. A multilingual relational lexical database of proper names, in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Marocco*, pp. 334–338.

Denis Maurel, Nathalie Friburger, Jean-Yves Antoine, Iris Eshkol-Taravella, and Damien Nouvel (2011), Cascades de transducteurs autour de la reconnaissance des entités nommées, *Traitement Automatiques des Langues*, 52(1):69–96.

Pablo Mendes, Max Jakob, and Christian Bizer (2012), DBpedia: A Multilingual Cross-domain Knowledge Base, in Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, European Language Resources Association (ELRA), Istanbul, Turkey.

George A. Miller (1995), WordNet: A Lexical Database for English, *Commun. ACM*, 38(11):39–41.

Hien Thang Nguyen and Tru Hoang Cao (2010), Enriching Ontologies for Named Entity Disambiguation, in *Proceedings of the 4th International Conference on Advances in Semantic Processing (SEMAPRO 2010)*, Florence, Italy.

Thanh Nguyen, Viviane Moreira, Huong Nguyen, Hoa Nguyen, and Juliana Freire (2011), Multilingual schema matching for Wikipedia infoboxes, *Proc. VLDB Endow.*, 5(2):133–144.

Georgios Petasis, Vangelis Karkaletsis, Georgios Paliouras, Anastasia Krithara, and Elias Zavitsanos (2011), Ontology Population and Enrichment: State of the Art, in *Knowledge-Driven Multimedia Information Extraction and Ontology Evolution*, volume 6050 of *Lecture Notes in Computer Science*, pp. 134–166, Springer.

Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors (2012), *Narodowy Korpus Języka Polskiego [Eng.: National Corpus of Polish]*, Wydawnictwo Naukowe PWN, Warsaw.

A.E. Richman and P. Schone (2008), Mining Wiki Resources for Multilingual Named Entity Recognition, in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 1–9.

Pavel Shvaiko and Jérôme Euzenat (2013), Ontology Matching: State of the Art and Future Challenges, *IEEE Trans. Knowl. Data Eng.*, 25(1):158–176.

Ralf Steinberger, Bruno Pouliquen, Mijail Alexandrov Kabadjov, Jenya Belyaeva, and Erik Van der Goot (2011), JRC-NAMES: A Freely Available, Highly Multilingual Named Entity Resource, in *Recent Advances in Natural Language Processing, RANLP 2011, 12-14 September, 2011, Hissar, Bulgaria*, pp. 104–110.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum (2007), YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia, in *WWW '07: Proceedings of the 16th International World Wide Web Conference*, pp. 697–706, Banff, Canada.

Antonio Toral, Sergio Ferrández, Monica Monachini, and Rafael Muñoz (2012), Web 2.0, Language Resources and standards to automatically build a multilingual Named Entity Lexicon, *Language Resources and Evaluation*, 46(3):383–419.

Antonio Toral, Rafael Muñoz, and Monica Monachini (2008), Named Entity WordNet, in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2008)*, European Language Resources Association, Marrakech, Morocco.

Mickael Tran and Denis Maurel (2006), Prolexbase: Un dictionnaire relatonnel multilingue de noms propres, *Traitement Automatiques des Langues*, 47(3):115–139.

Piek Vossen (1998), Introduction to EuroWordNet, *Computers and the Humanities*, 32(2-3):73–89.