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, Laboratoire d'informatique, France
[2] Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland
[3] Poleng, Poznań, Poland

ABSTRACT

Even if proper names play a central role in natural language processing (NLP) applications they are still under-represented 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 (NEs), 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 under-represented 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 Wikipedia and, to a lesser extent, from GeoNames.

Figure 1 shows the data flow in our Prolexbase population process. The 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, a candidate is automatically checked to see if it represents an entity which 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.

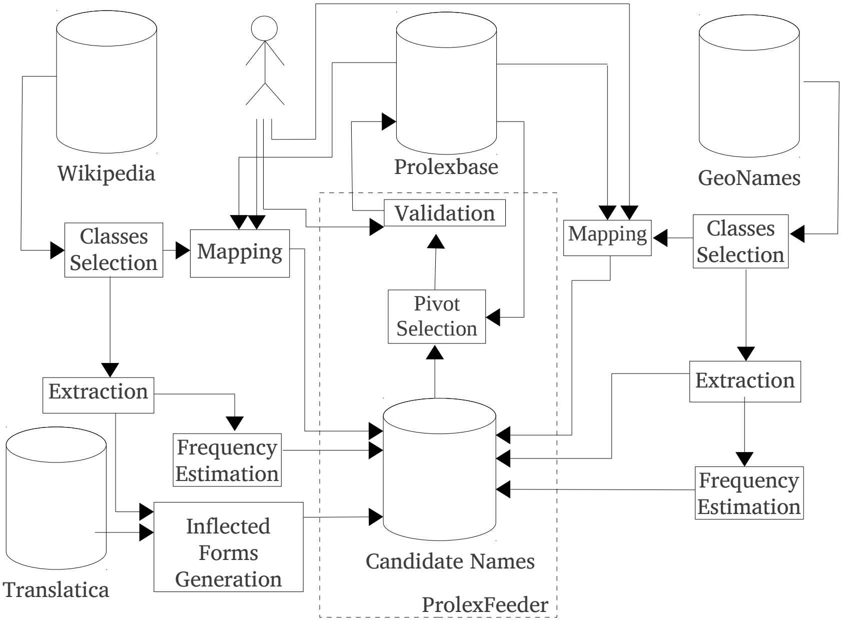

Figure 1: Data flow in Prolexbase population via ProlexFeeder.

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, which are 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.

Prolexbase might correspond to the *kernel NE lexicon*, i.e. the common shared NE vocabulary appearing in texts of differents dates, types and subjects, as opposed to the *peripheral NEs* used infrequently and in domain-specific jargons. As suggested by Saravanan *et al.* (2012), handling peripheral NEs might then rely on their co-occurence with the kernel NEs.

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 briefly address the human validation interface. Section 5 is dedicated to evaluation of the population process and of a named entity recognition system using the resulting Prolexbase resources. Section 6 contains a detailed discussion of related work. Finally, Section 7 concludes our contributions, and Section 8 summarizes some perspectives for future work and mentions possible applications of the rich Prolexbase model and data.

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 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*, understood as a collective anthroponym). **Accessibility** means that one referent is accessible through another, generally better known, referent (Tran and Maurel 2006). The accessibility **subject file** with 12 values (*relative*, *capital*, *leader*, *founder*, *follower*, *creator*, *manager*, *tenant*, *heir*, *headquarters*, *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

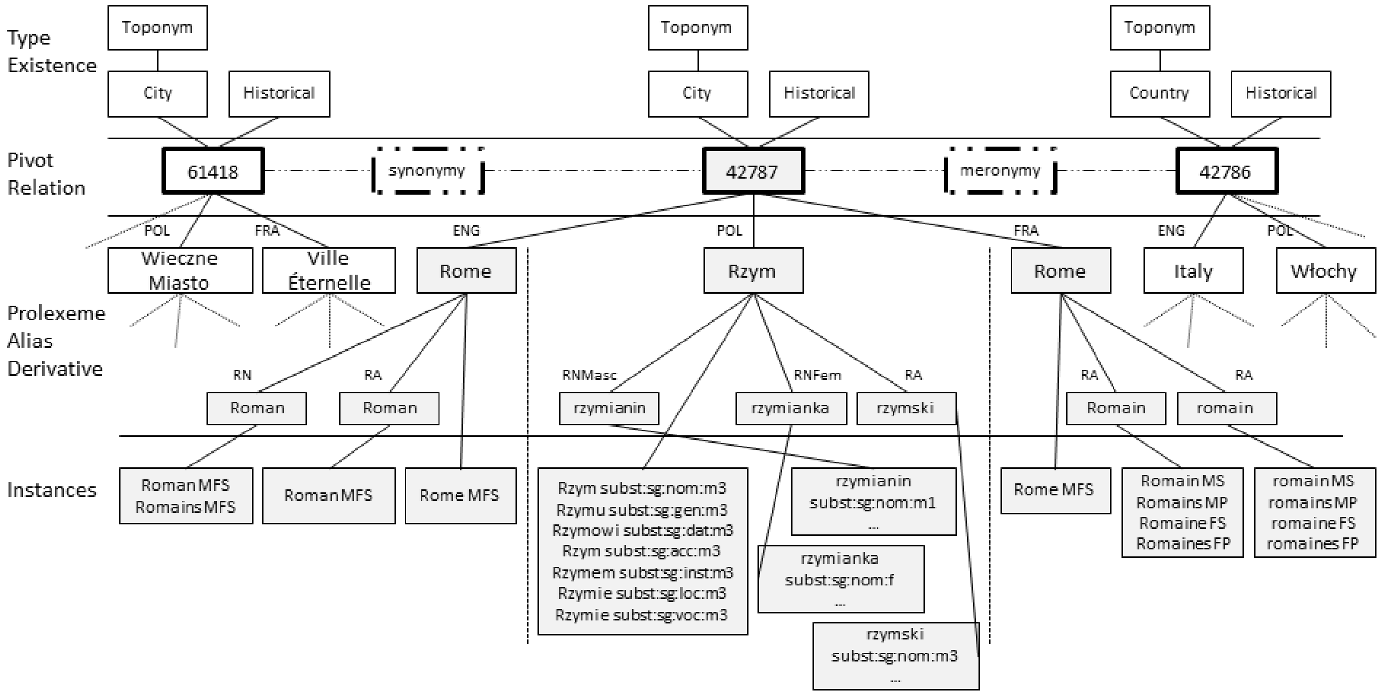

Figure 2: Extract of the intended contents of Prolexbase with four levels and three prolexemes.

a region) are represented by different prolexeme instances. 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 include, in particular: **classifying context** (the *Vistula river*), **accessibility context** (*Paris* – the *capital* of *France*), **frequency** (*commonly used*, *infrequently used* or *rarely used*), and **language** (association of each prolexeme to one 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.

Figure 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.

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, 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 open-source online encyclopaedia initiatives run by the MediaWiki software and filled with content by volunteer authors. Polish is the sixth largest

[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 *metaconcepts*, 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.

Wikipedia project with its 900,000 articles. We use the dump containing all Polish articles of the current release available at the beginning of 2011. 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.

Wikipedia 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.

Categories and *infobox templates* are two possible means of classifying Wikipedia articles. Both 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. Thus, 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 one to automatically access an article whose title is not identical with the query. Redirects may be used for various purposes including orthography and transcription variants, official names (1), eliptical variants, acronyms and abbreviations, diachronic variants (2), pseudonyms, common spelling errors and names with extra disambiguating data (3).

(1) Main Polish entry: *Wielka Brytania* 'Great Britain'
Redirects: *Zjednoczone Królestwo* 'United Kingdom', *Zjednoczo-*

ne Królestwo Wielkiej Brytanii i Irlandii Północnej ‘United Kingdom of Great Britain and Northern Ireland’

(2) Main Polish entry: *Plac Powstańców Warszawy w Warszawie* ‘Warsaw Uprising Square in Warsaw’
Redirects: *Plac Napoleona* ‘Napoleon Square’, *Plac Warecki* ‘Warka Square’

(3) Main English entry: *Sierra Blanca* (settlement in Texas)
Redirects: *Sierra Blanca (TX)*, *Sierra Blanca, TX*

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. Entries are categorized into 9 main classes which in turn divide into 645 highly fine-grained subcategories.[2] For instance, code *S.CAVE* refers to the subcategory *cave* of the main class *spot*. All the data are freely available under the Creative Commons Attribution license[3], 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.

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 contain explicit inflection or grammat-

[2] http://www.geonames.org/export/codes.html

[3] http://creativecommons.org/licenses/by/3.0/

ical 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, has a rich inflection system and instances have to be suggested automatically if the human validation is to be efficient. We use the inflection routine, based on dictionary lookup and guessing, developed for Translatica (Jassem 2004), a Polish-centred machine translation system. For over 260,000 extracted Wikipedia entries almost 2 million instances have been collected in this way. We used the *Morfologik* dictionary[4] as a source of inflected forms both for single entries and for components of multiword units. All Polish instances were further manually validated and corrected before their addition to Prolexbase (cf. Section 4).

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 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 from 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.

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

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 kompozytorzy 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 heterogeneous in nature. Orthography and transcription variants, official names (1), eliptical variants, acronyms and abbreviations represent aliases in terms of Prolexbase. Diachronic variants (2) and pseudonyms correspond to diachronic and diastratic synonyms, respectively. Spelling errors and variants with disambiguating data (3) are irrelevant. We could automatically remove only redirects of type (3), as well as those pointing at article subsections rather than articles themselves. The elimination of spelling errors, as well as the distinction between virtual aliases and synonyms had to be left for further manual validation stage (cf. Section 4). The final resulting collection 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 entries in GeoNames is huge it is hardly feasible to validate all of them manually before adding them to Prolexbase. Thus, it was necessary to select a well-defined subset of these data. We have used only the country names[5], all Polish names[6], as well as alternate names[7]. 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. Thus far, 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 ontology is mapped onto another and the granularity of both can be fully conserved.

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

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

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

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 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. Section 3.1.1) proved appropriate for a rather straightforward mapping onto Prolexbase types and existence values (historical, fictitious or religious). For instance, the Wikipedia infobox class *Postać telenowela* ('Soap opera character') could be mapped on Prolexbase type *Celebrity* and *fictitious* existence.

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 example, the Wikipedia category *Władcy Blois* ('Counts of Blois') was mapped on Prolexbase type *Celebrity*, *historical* existence, and *accessibility* relation with *Blois* with the *leader* subject file.

The mapping and the selection of related pivots was 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*.

3.3 *Frequency code estimation*

As mentioned in the Section 2.1, every language-specific entry (prolexeme) in Prolexbase obtains one of three *frequency* labels 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 based our estimation on monthly statistics of *Wikipedia hits*[8] from January 1st, 2010 to December 31st, 2010. We 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 rather arbitrarily[9] 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.

Note that these values are defined for prolexemes rather than pivots, e.g. a person may be very well known in Poland, thus it has frequency code 1 in Polish, while it gets code 2 or 3 in French or English.

The definition of frequency values for GeoNames followed the assumption that it was a secondary resource. Thus, each name appearing also in Wikipedia kept the Wikipedia hit-based frequency code. All other names of countries, European and Polish administrative division units, as well as capitals of these entities, were assigned code 1, since we wished to include these major classes on a systematic basis. The

[8] Available via the `http://stats.grok.se/` service

[9] A group of 3 French and Polish native experts examined the list of entries ordered according to the decreasing value of Wikipedia hits. The frequency code was supposed to be 1 as long as at least 2 entries known by at least one of the experts appeared in consecutive windows of about 30-entries. The threshold choice between code 2 and 3 was arbitrary.

remaining names were arbitrarily distributed over the 3 codes – see (Savary *et al.* 2013) for details.

3.4 *Pivot selection*

Data extracted from Wikipedia represent concepts and relations 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. 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 then 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 (1) ('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 Table 1 showing a sample set of Wikipedia data resulting (except the *pivot* attribute) from the preprocessing described in the preceding sections. Figure 3 sketches the algorithm of pivot selection for a new data set *e*. Its aim is to find each pivot *p* existing

Figure 3: Algorithm for selecting candidate pivots for a new incoming entry.

```
Function $getPivotCandidates(e)$ return $pivotList$
Input $e$: structure as in Table 1 //incoming entry
Output $pivotList$: ordered list of $(p,d)$ with $p,d \in \mathbb{N}$ //proposed pivots
and their distances from $e$
begin
  for each $l \in \{PL, EN, FR\}$ do
    $pivots.l \leftarrow \langle\rangle$ //empty list
  for each $p \in allPivots$ do //for each existing pivot
    for each $l \in \{PL, EN, FR\}$ do //for each language
      if $distance(e,p,l) < 10$ then
        $insertSorted(p, pivots.l)$ //insert the new pivot in the sorted
                                     candidates list
  //merge three sorted candidate lists into one
  $pivotList \leftarrow mergeSorted(pivots.PL, pivots.EN, pivots.FR)$
  if $pivotList = \langle\rangle$ then //no similar pivot found
    $pivotList \leftarrow \langle(getNewPivot(), 0)\rangle$ //create a new pivot
  return $pivotList$
end

Function $distance(e,p,l)$ return $d$
Input $e$: structure as in Figure 1 //incoming entry
      $p$: pivot
      $l \in \{PL, EN, FR\}$ //language
Output $d \in \{0,1,2,3,10\}$ //distance between $e$ and $p$
begin
  $d = 10$
  if $e.l.lex = p.l.lex$ then $d \leftarrow 0$ //same lexeme
  else if $e.l.lex \in p.l.aliases$ then $d \leftarrow 1$ //lexeme same as an alias
  else if $e.l.url = p.l.url$ then $d \leftarrow 2$ //matching Wiki URL
  if $d \leq 1$ and $e.l.url \neq p.l.url$ and $e.type \neq p.type$ then $d \leftarrow 3$
  return d
end
```

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. We assume that *e* is likely to be similar to *p* in any of the following cases: (i) *e* and *p* share the same lexeme in a particular language (line 15), (ii) *e* is an alias of *p* (line 16), (iii) *e* and *p* share the same URL (line 17). In the last case, a bi-directional matching of lexemes and aliases between Wikipedia and Prolexbase is not always a good strategy. For instance, the redirects in Example (2) 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 diachronic synonymy relation). Finally, we give a penalty if *e* shares the lexeme with the existing pivot *p* but either their URL or their type differ (line 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

Table 1: 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 Prolexbase type, meronymy-related pivot (*meroPivot*), existence and pivot.

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*	EN.aliases	*Capital of Italy,* *Castel Fusano,* *Città Eterna, …*	FR.aliases	*Ville Éternelle,* *Ville éternelle*
PL. infl	*Rzymu:sg:gen:m3,*	type	city	existence	historical
	Rzym:sg:acc:m3, …	meroPivot	none	pivot	42787

factors into account: (i) the rank of a pivot in each of the three lists, (ii) its membership in the intersections of these lists. If no similar pivot was found in any language then a new pivot is proposed (line 9–10).

The actual implementation of this algorithm does not scan all existing pivots for each incoming entry e. The entry is directly compared, instead, to the existing lexemes and aliases in the given language, which is optimal if data are indexed. For instance, if we admit that the database engine running Prolexbase implements indexes on B-trees, and that l, p and a denote the worst-case length of a candidate pivot list, the number of the existing prolexemes and of the existing aliases, respectively, the complexity of our algorithm is of $O(\log p + \log a + l)$. In practice, p was up to four times higher than a, and l was no higher than 10. The candidate pivot searching algorithm proved not to be a bottleneck of our procedure. On average, it takes less than a second to pre-process (off-line) a single Wikipedia entry.

The pivots returned by the algorithm in Figure 3 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. Table 1) 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 as 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 Figure 1 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 do 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 first views new entries proposed by the automatic selection and integration process then 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 which can be detected by the pivot selection procedure (cf. Section 3.4). In this case, the data coming from external sources complete those already present. Prolexemes in the three languages are proposed together with their Wikipedia URLs (which are usually new to Prolexbase). Some aliases, like *Wieczne Miasto* ('Eternal City') in Table 1, can be transformed into new pivots. Missing relations as well as derivations can be added manually, and the proposed inflected forms of the Polish prolexeme can be corrected or validated.

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. Table 2 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 contained many French toponyms, some celebrity names and only very few names of works. The true negatives correspond to the newly added 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.

Table 2: Results of ProlexFeeder on three sets of entries.

Type	Incoming entries	True posit.	True negat.	False posit.	False negat.	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

Wrongly selected pivots result mainly from the strict matching algorithm between an incoming lexeme and existing prolexemes and aliases (cf. Figure 3, 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 equivalents (*Muhammad Ali* and *Méhémet-Ali*) was not strong enough to allow for 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, the 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 until 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.

Prolexbase has already been successfully used for named entity recognition and categorization in French with an extended NE typology (Maurel *et al.* 2011). However, since 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[10], 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, NKJP (Przepiórkowski et al., 2012) divided into 10 parts of a roughly equal number of sentences. In each fold of the 10-fold cross validation Nerf was 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. We considered an NE as correctly recognized by Nerf if its span and type matched the reference corpus. In setting A the model obtained the mean F_1 measure of 0.76819 (with mean $P = 0.79325$ and $R = 0.74477$), while in setting B the mean F_1 measure was equal to 0.77409 (with mean $P = 0.79890$ and $R = 0.75092$). The paired Student's t-test yielded the p-value equal to 0.0001145 which indi-

[10] `http://zil.ipipan.waw.pl/Nerf`

cates that the results are statistically significant with respect to the the commonly used significance levels (0.05 or 0.01).

It should be noted that the majority of names appearing in the NKJP corpus correspond to person names, while Prolexbase contains a relatively small number of such names. Conversely, settlement names (cities, towns, villages, etc.) constitute a relatively high percentage of Prolexbase entries. In this subcategory the enhancement of Nerf's scores is the most significant – the mean F-measure increased by 0.03894 (from $F_1 = 0.79202$ to $F_1 = 0.83096$) and the Student's t-test p-value was equal to $8.011e{-}08$.

These results are encouraging, especially given the fact that Nerf's initial performances were rather good due to the big size and the high quality of the training corpus (NKJP), which had been annotated manually by two annotators in parallel, and then adjudicated by a third one.

6 RELATED WORK

Before ProlexFeeder was created, Prolexbase population had been performed mostly manually (Tran *et al.* 2005). Uniqueness of pivots was supported by a rather straightforward method based on a prolexeme match alone. Lists of entries and attributes were crafted in spreadsheet files which were then automatically inserted to Prolexbase provided that pivot identifiers appeared in them explicitly. Data were manually looked up in traditional dictionaries, lists and Internet sources. Inflected forms were generated via external tools. The complexity of the model hardly allowed the users to work in this way on more than one language or more than one type at a time. As a result, Prolexbase contained initially mainly French data. ProlexFeeder largely facilitates the lexicographer's work in that most data are automatically fetched, pivot uniqueness relies on more elaborate multilingual checks, entry validation is supported by automatic Prolexbase lookup, and inflected forms are automatically generated.

Prolexbase can be compared to *EuroWordNet* (EWN) (Vossen 1998) and to the *Universal Wordnet* (UWN) (Melo and Weikum 2009), although both of them are general-purpose wordnets with no particular interest in named entities. All three resources distinguish a language-independent and a language-specific layer. Language-independent entities, i.e. Interlingual Index Records (ILIRs) in EWN and pivots in Prolexbase, provide translation of lexemes in language-specific layers (but ILIRs unlike pivots form an unstructured list of meanings). UWN, conversely, provides direct translation links between terms in different languages. The main specificity of Prolexbase w.r.t. EWN and UWN is that proper names are concepts in Prolexbase while they are instances in EWN and UWN. Thus, adding a new proper name to Prolexbase implies enlarging its conceptual hierarchy, which does not seem possible e.g. with automatic UWN population algorithms.

Prolexbase population from Wikipedia and GeoNames can be seen as an instance of the *ontology learning* problem. According to the taxonomy proposed by Petasis *et al.* (2011), we simultaneously perform *ontology enrichment* (placing new conceptual proper names and relations at the correct positions in an existing ontology) and *ontology population* (adding new instances of existing concepts). The former is based on integrating existing ontologies (as opposed to constructing an ontology from scratch and specializing a generic ontology). The latter 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 fully manual mapping produces subsumption relations and results in a particular type of an n:1 alignment. Namely, a Wikipedia infobox class is mapped on one Prolexbase type and on a set of relations (cf. Section 3.2.1). 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). This mapping is inherently multilingual and subsumption-based. It outputs 1:n alignments, due to e.g. diachronic synonymy as in Example (2). It is supported by a full-fledged matching validation interface and leads to 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 was initially dedicated to English and is more recently being applied to other languages. Several approaches are based on aligning WordNet with Wikipedia: (Toral *et al.* 2008), (Toral *et al.* 2012), (Fernando and Stevenson 2012), (Nguyen and Cao 2010), *YAGO* (Suchanek *et al.* 2007) and *YAGO2* (Hoffart *et al.* 2011). Others build new semantic layers over Wikipedia alone: *Freebase* (Bollacker *et al.* 2007), *MENTA* (Melo and Weikum 2010), *DBpedia*[11] (Bizer *et al.* 2009; Mendes *et al.* 2012). DBpedia is the only resource to explicitly provide support for natural language processing tasks (data sets of variants, thematic contexts, and grammatical gender data).

Table 3 shows a contrastive study of these methods[12]. 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 three other approaches we consider 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 Universal

[11] `http://dbpedia.org`

[12] A more complete survey can be found in (Savary *et al.* 2013).

Wordnet but unlike others we order input entries by popularity (estimated via Wikipedia hits, while the UWN uses corpus frequencies). Like Freebase[13] but unlike 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,
- 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 one of the main sources of data for ontology creation and enrichment in the methods discussed above. An opposed point of view is represented within the Text Analysis Conference[14] (TAC) by the *Knowledge Base Population*[15] (KBP) task. In particular, its 2011 mono-lingual (English) and cross-lingual (Chinese-English[16])

[13] 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).

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

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

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

Table 3: 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

Entity Linking track is partly relevant to our work. In this track, the initial knowledge base (KB) consists of over 800,000 entities from English Wikipedia annotated with 4 types. Given a named entity and a source text in which it appears, the task is to provide the identifier of the same entity 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 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. We are aware of no experiments with an application of TAC-KBP-population methods to creating an actual mono- or multi-lingual lexical-semantic resource.

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 hundred thousand named entities, mainly person names, is being continuously developed for 20 languages. New names and their variants are extracted by a rule-based named-entity recognizer from 100,000 news articles per day and partly manually validated.

7 CONCLUSIONS

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 mainly from Wikipedia and partly from 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.

Table 4 shows the state of Prolexbase at the end of March 2013. 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 (48,340 pivots), celebrity (7,979 pivots), hydronym (4,580 pivots) and region (4,190 pivots), the number of pivots of the remaining types is between 1 and 1,374. 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 with meronymy and accessibility relations, respectively. With respect to the initial contents of Prolexbase, ProlexFeeder allowed us to add about 18,000 new pivots and 19,000 relations, as well as 23,000 Polish, 19,000 English and 15,000 French prolexemes. These new data required a manual workload of about 4 person-months.

Table 4: Current state of Prolexbase. Polish instances include inflected forms of prolexemes only

Pivots				
All	Toponyms	Anthroponyms	Ergonyms	Pragmonyms
73,405	81.3%	16.8%	1.4%	0.4%

Relations			
All	Meronymy	Accessibility	Synonymy
72,672	92.9%	5.3%	1.8%

	Pivots in synonymy relation	Pivots in meronymy relation	Pivots in acessibility relation
All	2,457 (3%)	65,768 (90%)	6,312 (9%)
Most frequent types	celebrity 1,325 (17%)	city 48,110 (100%)	city 2,214 (5%)
	country 390 (45%)	celebrity 7,053 (88%)	region 1,696 (40%)
	city 157 (0.3%)	region 4,052 (97%)	celebrity 1,129 (14%)

Language	Prolexemes	Aliases	Derivatives	Instances
PL	27,408	8,724	3,083	166,479
EN	19,492	14,039	94	18,575
FR	70,869	8,488	20,919	142,506

The Prolexbase data are referenced in the META-SHARE infrastructure[17] and available[18] under the CC BY-SA license[19], i.e. the same as for Wikipedia and GeoNames. We are currently working on their LMF exchange format according to Bouchou and Maurel (2008).

8 PERSPECTIVES

Prolexbase is an open-ended project. Many perspectives exist for Prolexbase itself, for the ProlexFeeder functionalities, and for future applications exploiting the rich Prolexbase model.

[17] http://www.meta-net.eu/meta-share

[18] Downloadable from http://zil.ipipan.waw.pl/Prolexbase

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

8.1 Data and model evolutions

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 evolve, 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 its corpus frequency, provided that some word sense disambiguation techniques are available.

Note also that only a part of the relations modelled in Prolexbase has been actually dealt with in ProlexFeeder. The remaining linguistic-level relations, notably classifying contexts, are still to be described. Pragmonyms and ergonyms are under-represented 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*[20].

In order to ensure an even better pivot selection process, matching prolexemes and aliases could be enhanced by approximate string matching and other methods used in related work. 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. How-

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

ever, some types are concerned with 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 (Przepiórkowski *et al.* 2012). Human or automatic annotation of such a corpus can greatly benefit from a rich linguistic resource 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.

8.2 *Future Applications*

Named entity recognition tools such as Nerf (cf. Section 5) do not yet manage to fully exploit the richness of an advanced annotation schema like that of the National Corpus of Polish. In particular they currently fail to provide lemmas for the recognized NEs and derivational bases (*Wielka Brytania* ’Great Britain’) for the relational adjectives (*angielski* ’English’) and inhabitant names (*Anglik* ’Englishman’). Prolexbase relations will allow NER tools to bridge this gap, by offering explicit links between different inflectional and derivational variants of proper names and their base prolexemes. They may also serve as a training material for establishing lemmas and derivational bases for less popular proper names.

Other possible applications of Prolexbase are to be seen in establishing relations between named entities in corpora. Note that the synonymy between pivots, as well as all lexical relations among prolexemes and instances, allow us to straightforwardly link variants of a proper name, thus providing a reliable resource for coreference resolution. Furthermore, the meronymy and accessibility relations constitute a means of finding and labeling bridging (associative) anaphora.

Prolexbase is now also a good candidate to experiment with an advanced version of the Entity Linking process (cf. Section 6). Instead of linking NEs occurrences to Wikipedia entries we might map them on Prolexbase, which offers a pure taxonomy and an elaborate set of manually validated relations. Thus, we would obtain a high quality Word Sense Disambiguation (Fernando and Stevenson 2012) resource for “kernel” NEs.

The most elaborate use of the fine-grained Prolexbase model is expected in the domain of machine translation of proper names (Graliński *et al.* 2009). The original idea of a conceptual proper name being a pair of a referred object and a point of view on this object allows 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 or inexistent in others, relations like the classifying context or the accessibility context enable explanation-based translations (e.g. *Hanna Gronkiewicz-Waltz ⇒ Hanna Gronkiewicz-Waltz, the president of Warsaw, blésois ⇒ an inhabitant of Blois*).

Other potential applications include: (i) multilingual named entity recognition (NER) (Richman and Schone 2008), (ii) text classification (Kumaran and Allan 2004), (iii) uni- or cross-lingual question answering (Ferrández *et al.* 2007), and (iv) proper name normalization (Jijkoun *et al.* 2008).

9 ACKNOWLEDGEMENTS

We are deeply indebted to three anonymous reviewers for their insightful critical remarks and advice on previous versions of our paper.

They allowed us to increase its quality and to discover new perspectives for future work. We are also grateful to Denis Maurel for having shared his expertise on the Prolexbase ontology, as well as to Jakub Waszczuk for the integration of Prolexbase data in the Nerf system and providing experimental results for named entity recognition in Polish.

This work has been carried out within two projects: (i) *Nekst*[21], funded by the European Regional Development Fund and the Polish Ministry of Science and Higher Education, (ii) CESAR[22] – a European project (CIP-ICT-PSP-271022), part of META-NET.

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.

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*, p. 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)*, p. 88–96, Barcelona.

[21] `http://www.ipipan.waw.pl/nekst/`

[22] `http://www.meta-net.eu/projects/cesar`

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)*, p. 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*, p. 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, p. 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*, p. 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, p. 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*, p. 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.

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*, p. 1099–1108, ACM.

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.

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*, p. 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.

Alexander E. RICHMAN and Patrick SCHONE (2008), Mining Wiki Resources for Multilingual Named Entity Recognition, in Kathleen MCKEOWN, Johanna D. MOORE, Simone TEUFEL, James ALLAN, and Sadaoki FURUI, editors, *ACL*, pp. 1–9, The Association for Computer Linguistics, ISBN 978-1-932432-04-6.

K. SARAVANAN, Monojit CHOUDHURY, Raghavendra UDUPA, and A. KUMARAN (2012), An Empirical Study of the Occurrence and Co-Occurrence of Named Entities in Natural Language Corpora, 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.

Agata SAVARY, Leszek MANICKI, and Małgorzata BARON (2013), ProlexFeeder – Populating a Multilingual Ontology of Proper Names from Open Sources, Technical Report 306, Laboratoire d'informatique, François Rabelais University of Tours, France.

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*, p. 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*, p. 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.

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

Mickaël TRAN, Denis MAUREL, Duško VITAS, and Cvetana KRSTEV (2005), A French-Serbian Web Collaborative Work on a Multilingual Dictionary of Proper Names, in *Proceedings of the 6th Workshop on Multilingual Lexical Databases (PAPILLON'05), Chiang Rai, Thailand.*

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

This work is licensed under the Creative Commons Attribution 3.0 Unported License.
`http://creativecommons.org/licenses/by/3.0/`