LMF Format for Polish Named Entity Gazetteer version 1.3

Agata Savary,
Université François Rabelais Tours,

February, 2012

This document addresses the definition of a standard data exchange format for a Polish linguistic resource dedicated to named entities (NEs). The format is meant to follow the recommendations of the Lexical Markup Framework (LMF).

1 Polish named entity gazetteer

The Polish Named Entity Gazetteer (PNEG) is a textual resource used within the *SProUT* platform [1], initially for information extraction from Polish texts [3], and then, in its extended version, for the automatic pre-annotation of the National Corpus of Polish (NKJP) on the level of named entities. It is available at the CLIP web page (http://clip.ipipan.waw.pl/) under the 2-clause BSD license. Its construction, contents and use have been described in [5] and [4].

The file contains 153,477 inflected entries of Polish (and some foreign) proper names and named entity components:

- forenames and surnames,
- city, country, mountain, region and river names,
- institution names,
- relational adjectives and inhabitant names stemming from country names[1],
- named entity triggers (months, days, positions, etc.).

Each line in the gazetteer describes an inflected form of a Polish named entity. The inflected form itself appears at the beginning of the line and is accompanied by a list of attribute-value pairs expressing grammatical and semantic features as well as metadata. The list of possible attributes and their values is determined by the type hierarchy defined for *Sprout* (see [5]). None of the attributes is compulsory and the order of attributes in an entry is arbitrary. Examples (1)–(13) show some gazetteer entries:

(1) **Bug** | GTYPE:gaz_river | G_LEMMA:Bug | G_NUMBER:singular | G_CASE:nom | G_GENDER:masc3 | G_SOURCE:Wikipedia | G_INFL_SOURCE:Morfeusz

[1] PNEG does not contain inhabitant names and relational adjectives stemming from Polish settlements. These data, owned by the PWN publisher, were used within the NKJP project under a particular license and are concerned by the copyright.

(2) **Buga** | GTYPE:gaz_river | G_LEMMA:Bug | G_NUMBER:singular | G_CASE:gen | G_GENDER:masc3 | G_SOURCE:Wikipedia | G_INFL_SOURCE:Morfeusz

(3) **Kowalskim** | GTYPE:gaz_surname | G_LEMMA:Kowalski | G_NUMBER:singular | G_CASE:ins | G_GENDER:masc1 | G_SOURCE:KubaW | G_INFL_SOURCE:Morfeusz

(4) **Janie** | GTYPE:gaz_given_name | G_LEMMA:Jan | G_CASE:loc_voc | G_GENDER:masc1 | G_SOURCE:old-gaz-2

(5) **NIK** | GTYPE:gaz_institution | G_LEMMA:NIK | G_COUNTRY:Polska | G_SUBTYPE:national | G_FULL_FORM:Najwyższa Izba Kontroli | G_SOURCE:old_gaz-1

(6) **Polski** | GTYPE:gaz_country | G_LEMMA:Polska | G_NUMBER:singular | G_CASE:loc | G_GENDER:fem| G_SOURCE:KSNG | G_INFL_SOURCE:Morfeusz

(7) **polskie** | GTYPE:gaz_country_deriv | G_LEMMA:polski | G_NUMBER:singular | G_CASE:nom_voc | G_GENDER:neutrum1_neutrum2 | G_DERIV_TYPE:reladj | G_DERIVED_FROM:Polska | G_SOURCE:Wikislownik | G_INFL_SOURCE:Morfeusz

(8) **Polakowi** | GTYPE:gaz_country_deriv | G_LEMMA:Polak | G_NUMBER:singular | G_CASE:dat | G_GENDER:masc1 | G_DERIV_TYPE:persderiv | G_DERIVED_FROM:Polska | G_SOURCE:Wikislownik | G_INFL_SOURCE:Morfeusz

(9) **Polaki** | GTYPE:gaz_country_deriv | G_LEMMA:Polak | G_REGISTER:depr | G_NUMBER:plural | G_CASE:nom | G_GENDER:masc2 | G_DERIV_TYPE:persderiv | G_DERIVED_FROM:Polska | G_SOURCE:Wikislownik | G_INFL_SOURCE:Morfeusz

(10) **inspektorowi** | GTYPE:gaz_position | G_LEMMA:inspektor | G_CASE:dat | G_GENDER:masc1 | G_NUMBER:singular | G_SOURCE:old-gaz-2

(11) **kwietnia** | GTYPE:gaz_month | G_LEMMA:kwiecień | G_NUM_BASE:04 | G_NUMBER:singular | G_CASE:gen | G_GENDER:masc3 | G_INFL_SOURCE:Morfeusz

(12) **4** | GTYPE:gaz_dofm | G_LEMMA:04 | G_NUM_BASE:04 | G_SOURCE:old-gaz-1

(13) **04** | GTYPE:gaz_second | G_LEMMA:04 | G_NUM_BASE:04 | G_SOURCE:old-gaz-1

Examples (2)–(6) describe genuine proper names while examples (7)–(9) show a relative adjective and two personal derivatives stemming from a country name. Examples (10)–(13) correspond to trigger words (so-called internal and external evidences) used in grammar rules to describe contexts of NE occurences.

The inflected forms may be both single words and compounds as in examples (14)–(18):

(14) **Bośni i Hercegowiny** | GTYPE:gaz_country | G_LEMMA:Bośnia i Hercegowina | G_NUMBER:singular | G_CASE:gen | G_GENDER:fem | G_SOURCE:KSNG | G_INFL_SOURCE:Morfeusz_Multiflex

(15) **Skarżyskiem-Kamienną** | GTYPE:gaz_city | G_LEMMA:Skarżysko-Kamienna | G_NUMBER:singular | G_CASE:ins | G_GENDER:neutrum2 | G_SOURCE:WorldGazetteer | G_INFL_SOURCE:Morfeusz_Multiflex

(16) **Wybrzeża Lazurowego** | GTYPE:gaz_region | G_CASE:gen | G_COUNTRY:Francja | G_LEMMA:Wybrzeże_Lazurowe | G_SOURCE:old-gaz-1

(17) **Irlandia Północna** | GTYPE:gaz_region | G_CASE:nom | G_COUNTRY:Wielka_Brytania | G_LEMMA:Irlandia_Północna | G_SOURCE:old-gaz-1

(18) **Najwyższa Izba Kontroli** | GTYPE:gaz_institution | G_LEMMA:Najwyższa Izba Kontroli | G_COUNTRY:Polska | G_SUBTYPE:national | G_SOURCE:old_gaz-1

1.1 Semantic attributes

The semantic information is carried in the gazetteer by five attributes:

- GTYPE indicates the NE's category; it can take any value of the following flat (i.e. not hierarchically organised) list:
  - gaz_numberword (e.g. *miliardy*)
  - gaz_position (e.g. *senatorka*)
  - gaz_name_infix (e.g. *van der*).
  - gaz_initial (e.g. *A*).
  - gaz_title (e.g. *prof. dr.*).
  - gaz_given_name (e.g. *Zygmunt*).
  - gaz_surname (e.g. *Kowalski*).
  - gaz_city (e.g. *Żytomierz*).
  - gaz_country (e.g. *Zimbabwe*).
  - gaz_region (e.g. *Szkocja*)
  - gaz_mountains (e.g. *Sudety*)
  - gaz_river (e.g. *Bug*)
  - gaz_continent (e.g. *Afryka*)
  - gaz_second (e.g. *59*)
  - gaz_minute (e.g. *59*)
  - gaz_hour (e.g. *23*)
  - gaz_dofm – day of month in numerical form (e.g. *31*)
  - gaz_dofw – day of week (e.g. *poniedzialek*)
  - gaz_month_num – month in numerical form (e.g. *01*)
  - gaz_month – month (e.g. *styczeń*)
  - gaz_year (e.g. *2010*)
  - gaz_year_short (e.g. *10*)
  - gaz_committee (e.g. *Host Committee*)
  - gaz_company (e.g. *Żywiec*)
  - gaz_institution (e.g. *Europejski Bank Centralny*)
  - gaz_university (e.g. *Heriot-Watt University*)
  - gaz_country_deriv (e.g. *polski, Polak, Polka*)
- G_SUBTYPE gives a secondary type for 312 organisation names, as in example (5). It can be an unrestricted string. At present, the following list of values is used: *bank, brokerage_house, insurance, investment_fund, joint_stock_company, national, national_investment_fund, pension_fund.*
- G_COUNTRY describes the meronymy relation between organisations (companies, institutions and universites) as well as regions, and countries they are located in, as in example (16).

- G_CITY has a similar meaning as G_COUNTRY but concerns only two entries (cf. *WSP w Częstochowie*)
- G_FULL_FORM fives the full form of an acronym, as in example (5).
- G_DERIV_TYPE is reserved to derivatives stemming from proper names and is equal to *reladj* for relational adjectives (e.g. *polski*) and to *persderiv* for inhabitant names (e.g. *Polka*).
- G_DERIVED_FROM comes with the previous attribute and shows the proper name the derivative is (semantically and not necessarily morphologically) derived from (e.g. *Polska*).

It would be useful to automatically check the completeness of the above lists of semantic attributes and their values in the PNEG entries.

1.2 Grammatical Attributes

Five attributes describe the grammatical properties of the gazetteer entries:

- G_LEMMA indicates the grammatical lemma of the given entry. It is an unrestricted string. The methodology of determining its value has changed in time and thus some inconsistencies can be found. For instance, in the older version of the gazetteer (see entries marked by G_SOURCE:old_gaz-1) in each multi-word lemma the blanc spaces are replaced by underscores as in example (16). These underscores should be omitted in order to obtain the fully correct lemma.
- G_CASE indicates the grammatical case of the inflected forms. It can take:
  - a single value from the following list: *nom, gen, dat, acc, ins, loc, voc*
  - a value resulting from merging several values above (in alphabetical order) as in example (4). At present, the full list of such combined values in the gazetteer is as follows: *acc_dat_gen_ins_loc_nom_voc, acc_gen, acc_ins, acc_nom, acc_nom_voc, dat_gen, dat_gen_loc, dat_gen_voc, dat_loc, gen_dat_loc, gen_loc_voc, ins_loc, loc_voc, nom_voc.* Some erroneous values like *acc_dac_gen_ins_loc_nom_voc, acc_gen_dat_loc_ins_nom_voc, nom_voc_* appear for a few entries. They should either be corrected or not taken into account.
- G_NUMBER indicates the grammatical number of the inflected forms. It can take any of the two values: *singular* or *plural.* Several entries contain the erroneous value *singula.* It should either be corrected or not taken into account.
- G_GENDER indicates the grammatical gender of the inflected forms. It can take:
  - a single value from the following list: *masc1, masc2, masc3, fem, neutrum1, neutrum2, plnum1, plnum2, plnum3*

  - a value resulting from merging several values above (in alphabetical order). At present, the full list of such combined values in the gazetteer is as follows: *fem_masc1_masc2_masc3_neutrum1_neutrum2_plnum1_plnum2_plnum3, fem_masc2_masc3_neutrum1_neutrum2_plnum2_plnum3, masc1_masc2, masc1_masc2_masc3, masc1_masc2_masc3_neutrum1_neutrum2, masc1_plnum1, neutrum1_neutrum2.*

- G_REGISTER indicates that the corresponding noun is depreciative (*depr*). It appears in the gazetteer only with surnames (e.g. *Ambroziaki* instead of *Ambroziakowie*) and derivatives from country names, as in example (9), in nominative and vocative gender, plural number and masculine animate gender.

1.3 Metadata

Most entries contain one or both of the following attributes:

- G_SOURCE indicates the source of the entry. Its value can be an unrestricted text. At present, the full list of values is as follows:
  - *KSNG* – stands for *Komisja Standaryzacji Nazw Geograficznych*, a Polish institution for standardization of geographical names,
  - *Wikipedia* – stands for Wikipedia (`http://pl.wikipedia.org`),
  - *Wikislownik* – stands for Wikisłownik (`http://pl.wiktionary.org`),
  - *WorldGazetteer* – stands for the World Gazetteer (`http://world-gazetteer.com`),
  - *KubaW* – stands for *Kuba Waszczuk* who provided a list of Polish family names found in the Internet (`http://www.futrega.org/etc/nazwiska.html`),
  - *old_gaz-1*, *old-gaz-1* and *old-gaz-2* – stand for the original gazetteers constructed by [3].
- G_INFL_SOURCE indicates the source of the inflection paradigm of the entry's lemma. At present, the full list of values is as follows:
  - *Morfeusz* – the entry has been produced by inflecting its lemma with *Morfeusz SGJP* [6]. One entry contains an erroneous value „*Mp orfeusz*". It should either be corrected or not taken into account.
  - *Multiflex_Morfeusz* – the entry is a multi-word unit and has been produced by inflecting its lemma with *Morfeusz SGJP* for single components, and with *Multiflex* for the whole compound [6].
  - *Recznie* – the entry has been produced manually.

1.4 Other attributes

Some entries describing given names and country-based relative adjectives, especially from the older versions of the PNEG, are duplicated for different lower case or upper case spelling variants. The attribute G_LETTER_CASE is used to indicate the variant type as in examples (19)–(20) vs. examples (4) and (7). Since the letter case variants can be generated quite straightforwardly from the canonical lemmas I suggest not to export entries having the G_LETTER_CASE attribute.

(19) **JANIE** | GTYPE:gaz_given_name | G_LEMMA:Jan | G_CASE:loc_voc | G_LETTER_CASE:**all-upper** | G_GENDER:masc1 | G_SOURCE:old-gaz-2

(20) **Polskie** | GTYPE:gaz_country_deriv | G_LEMMA:polski | G_NUMBER:singular | G_CASE:nom_voc | G_GENDER:neutrum1_neutrum2 | G_DERIV_TYPE:reladj | G_DERIVED_FROM:Polska | G_SOURCE:Wikislownik | G_INFL_SOURCE:Morfeusz | G_LETTER_CASE:**first-upper**

2 Designing an LMF format for PNEG

2.1 Lexical Markup Framework

Lexical Markup Framework (LMF), as defined in Wikipedia[2] is *the ISO International Organization for Standardization ISO/TC37 standard for natural language processing (NLP) and machine-readable dictionary (MRD) lexicons. The goals of LMF are to provide a common model for the creation and use of lexical resources, to manage the exchange of data between and among these resources, and to enable the merging of large number of individual electronic resources to form extensive global electronic resources. [...]*

The LMF documentation[3] specifies that *LMF models are represented by **UML classes**, associations among the classes, and a set of ISO 12620 **data categories** that function as UML attribute-value pairs. [...] Lexicon developers shall use the classes that are specified in the LMF **core package** [...]. Additionally, developers can optionally use classes that are defined in the **LMF extensions**. Developers shall define a **data category selection (DCS)** as specified for LMF data category selection procedures.*

2.2 Merging Forms into Lexemes

The LMF core package is shown in Fig.1. A ⟨`LexicalResource`⟩ is associated with ⟨`GlobalInformation`⟩ representing the data concerning the entire resource such as the `/languageCoding/`. A ⟨`LexicalResource`⟩ contains one or more ⟨`Lexicon`⟩s. In our case, it remains to be determined (see below) if only one ⟨`Lexicon`⟩ should be created or if the PNEG entries should be divided into several lexicons associated to several gazetteer versions identifiable by the G_SOURCE attribute.

From the point of view of the PNEG resource it is important to note that an LMF representation of a lexicon is lexeme-oriented while the PNEG is form-oriented. In other words, in LMF, one element of an LMF ⟨`Lexicon`⟩ is a ⟨`LexicalEntry`⟩ i.e. a lexeme with its inflected forms and variants and with its semantic information. Conversely, one entry in PNEG is an inflected form of a lexeme, thus different inflected forms of the same lexeme are spread through different lexicon entries, as seen in examples (1)–(2). Merging these separate PNEG entries into one LMF ⟨`LexicalEntry`⟩ might be a non trivial issue due to redundancy and homonymy.

[2] `http://en.wikipedia.org/wiki/Lexical_markup_framework`

[3] `http://www.tagmatica.fr/lmf/iso_tc37_sc4_n453_rev16_FDIS_24613_LMF.pdf`

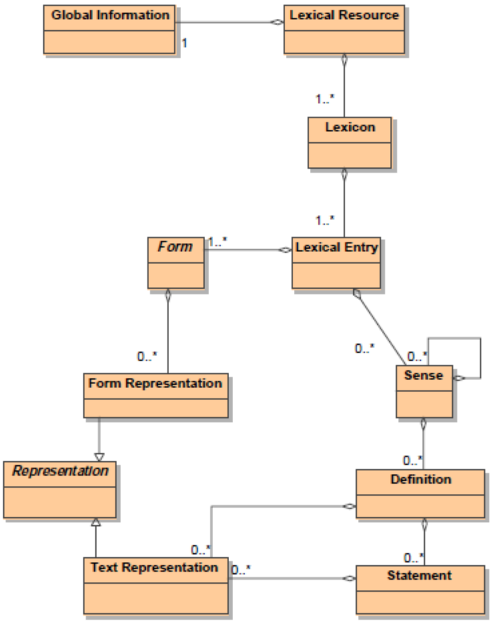

Figure 1: LMF Core Package

2.2.1 Redundancy

Examples (21)–(22) show two PNEG entries which are redundant due to the fact that earlier versions of the gazetteer were merged with newly created resources.

(21) **Kowalski** | GTYPE:gaz_surname | G_LEMMA:Kowalski | G_NUMBER:singular | G_CASE:nom | G_GENDER:masc1 | G_SOURCE:KubaW | G_INFL_SOURCE:Morfeusz

(22) **Kowalski** | GTYPE:gaz_surname | G_LEMMA:Kowalski | G_SOURCE:old-gaz-1

Reasonable heuristics would be to eliminate a redundant entry $E1$ if its set of attribute-value pairs (except G_SOURCE and G_INFL_SOURCE) is a subset of those in $E2$. It remains however to be checked if this rule should not be more fine-grained.

2.2.2 Homonymy

Dividing PNEG into two or more distinct resources would not however solve the problem of homonymy. Examples (23)–(24) show two PNEG entries with the same lemma and inflected forms but belonging to different lexemes.

(23) **Kuba** | GTYPE:gaz_country | G_LEMMA:Kuba | G_NUMBER:singular | G_CASE:nom | G_GENDER:fem | G_SOURCE:KSNG | G_INFL_SOURCE:Morfeusz

(24) **Kuba** | GTYPE:gaz_surname | G_LEMMA:Kuba | G_NUMBER:singular | G_CASE:nom | G_GENDER:fem | G_SOURCE:KubaW | G_INFL_SOURCE:Morfeusz

It seems that a good strategy would be to merge two entries into one lexeme only if their attributes, except G_SOURCE and G_INFL_SOURCE , are not in a conflict, i.e. each attribute either appears in one entry only or in both with the same value. The resulting entry is a fusion of attributes, while the values of G_SOURCE and G_INFL_SOURCE are taken from the more precise entry. Entry $E1$ is more precise than $E2$ if the number of attributes (except G_SOURCE and G_INFL_SOURCE) in $E1$ is higher than the one in $E2$. If both numbers are equal than the preferred values for G_SOURCE and G_INFL_SOURCE are different from *unknown*, whenever possible.

2.3 Splitting Forms into Sets of Forms

As discussed in section 1.2, earlier versions of PNEG (marked by `old-gaz-1`, `old_gaz-1` or `old-gaz-2`) use joint grammatical labels like e.g. `gen_loc_voc`. Each entry containing such a label should be split in the LMF lexicon into as many word forms as there are possible combinations of values resulting from splitting all labels in this entry. For instance the entry (7) should be split into 4 inflected forms belonging to the same lexeme.

2.4 Splitting Forms into Complete and Incomplete Ones

Let's reconsired the conditions of eliminating redundant entries described in section 2.2.1. Note that many entries stemming from the older gazetteer versions and having incomplete lists of attributes do not have more complete equivalents in the newer gazetteer version, as in example (25).

(25) **Abruzją** | GTYPE:gaz_region | G_CASE:ins | G_COUNTRY:Włochy | G_LEMMA:Arbuzja | G_SOURCE:old-gaz-1

Such an entry lacks in particular the necessary grammatical features of gender and number. Therefore, we suggest to divide the gazetteer's LMF version into two subsets:

- The first one would contain the conversion of only those entries whose grammatical features are complete. Namely a noun and an adjective must have a value for G_CASE, G_NUMBER and G_GENDER.
- The other one would contain all other, i.e. grammatically incomplete extries (after having initially eliminated or merged those of them to which the rules in sections 2.2.1 and 2.2.2 apply).

In particular, the gazetteer contains a set of entries composed only of digits, as in examples (12)–(13). They can appear in dates and other numerical expressions and have the attribute GTYPE equal to one of the following values: *gaz_second* (e.g. *59*), *gaz_minute* (e.g. *59*), *gaz_hour* (e.g. *23*), *gaz_dofm* (e.g. *31*), *gaz_month_num* (e.g. *01*), *gaz_year* (e.g. *2010*) or *gaz_year_short* (e.g. *10*). Formally, most of such entries can be seen either as adjectives (e.g. *01* pronunced *pierwszy*) or as numerals (e.g. *01* pronunced *jeden*), thus they should have the necessary attributes of number, gender and case. Since such attributes are missing, these entries should be contained in the LMF subset of incomplete entries.

2.5 Representing Lexemes

As soon as the (possibly split) PNEG entries have been merged into lexemes, each lexeme is to be represented according to the core package shown in Fig.1, together with its ⟨`Form`⟩s and its ⟨`Sense`⟩ description.

2.5.1 Lemmas and Word Forms

The inflection paradigm of each lexeme can be represented in LMF either extensionally (by listing all inflected forms) or intentionally (by describing morphological patterns governing the lexeme's inflection). Clearly, PNEG contains data of an extensional nature, and so will its LMF version if it is to be obtained most straightforwardly. Mechanisms supporting extensional lexicons are provided in LMF by its **morphology extension**. According to its specifications, each ⟨`LexicalEntry`⟩ contains one canonical form designed as ⟨`Lemma`⟩ and none or more ⟨`Word Form`⟩s. Both ⟨`Lemma`⟩ and ⟨`Word Form`⟩ are subclasses of the ⟨`Form`⟩ class.

All high level structural elements defined in LMF, such as ⟨`Lemma`⟩ and ⟨`Word Form`⟩, are adorned by low level standardized constants called **data categories** (DTs). In other words, each ⟨`Lemma`⟩ and ⟨`Word Form`⟩ elements are described by feature structures in which feature names stem from the *Data Category Registry (ISOcat)*[4], while each feature value either stems from ISOcat or is an unrestricted string.

The Polish tagset in its National Corpus of Polish version has been defined within ISOcat [2]. It is however non conformant with the tagset stemming from Morfeusz used in the gazetteer. Polish tagset conversion is a non-trivial problem in the general case, however named entities are limited to parts of speech (mainly nouns nad adjectives) whose grammatical features seem to be convertible in a rather straightforward manner. Namely:

- All adjectival entries (e.g. with attribute G_DERIV_TYPE equal to *reladj*) having the attribute G_GENDER equal to *neutrum1*, *plnum1*, *plnum2* or *plnum3* are to be eliminated[5].
- **Case** values are to be recopied from the G_CASE attribute except *ins* which gets replaced by *inst.*
- **Number** values stemming from the G_NUMBER attribute are to be systematically relabelled: *singular* and *singula* (see section 1.2) into *sg* and *plural* into *pl.*
- **Gender** values stemming from the G_GENDER attribute are to be limited into a narrower set containing *m1*, *m2*, *m3*, *f* and *n*, according to the following rules:
  - values *masc1*, *masc2*, *masc3*, *fem* and *neutrum2* are to be relabelled into *m1*, *m2*, *m3*, *f* and *n*, respectively,

[4] See http://www.isocat.org/. The data category constants in this interface are browsable by their "human" name while their precise values are those marqued as `identifier` in the Administration Information Section. For each feature, its possible values are listed in the Conceptual Domain section.

[5] That is because the same forms appear already with G_GENDER equal to *neutrum2*.

  - values *plnum2* and *plnum3* (for nouns only, since the adjectives with these genders have been eliminated in the first step) are to be relabelled into *n*,
  - values *neutrum1* and *plnum1* never appear with nominal entries in the gazetteer (and adjectives with these genders have been eliminated in the first step), thus no conversion is needed for these values.

- **Depreciativity** value *depr* becomes a separate part of speech, i.e. all forms of a lexeme $L1$ that have this attribute (there are always at most 2 such forms in plural, nominative and vocative, masculine human gender) are eliminated from $L1$ and create a new lexeme $L2$ such that: (i) the lemma of $L2$ is identical as in $L1$, (ii) the *partOfSpeech* of $L2$ is equal to *depr*, (iii) the sense of $l2$ is necessarily different than in $L1$ because LMF does not allow the same sense for two different lexemes, (iv) the sense of $L2$ is linked with the sense of $L1$ by a relation called *depreciativeVariant* (see section 2.5.3). Note in particular that the set of word forms of $L2$ will not contain its own lemma (which is in singular).

For instance, the grammatical attributes in example (2) can be expressed by the XML feature structure instantiations in Fig. 2.

```
<Lemma>
   <feat att="writtenForm" val="Bug">
</Lemma>
<WordForm>
   <feat att="writtenForm" val="Buga"/>
   <feat att="number" val="sg"/>
   <feat att="gender" val="m1"/>
   <feat att="case" val="gen"/>
</WordForm>
```

Figure 2: Expressing the grammatical features of the word form *Buga*.

2.5.2 Parts of Speech and Metadata

It is useful to represent the part of speech of each lexeme, and a relevant data category (`partOfSpeech`) is provided in the *Morphosyntax/PartOfSpeech* (but not in the *NKJP*) ISOcat directory. The *NKJP* directory on its turn offers standard names for this attribute's values compliant with the NKJP tagset. No corresponding attribute is however explicitly provided in the PNEG entries. The part of speech can be deduced though by examining the presence of the G_DERIV_TYPE attribute in a PNEG entry, as well as the form itself. If this attribute is present and has value *reladj* the corresponding entry is an adjective. Otherwise, if the form consists only of digits (e.g. *1980*) than it is both an adjective and a numeral. Otherwise, most probably it is a noun.

As mentioned before, most PNEG entries contain attributes showing the provenance of the lemma (G_SOURCE) and/or of its inflected form (G_INFLECTION_SOURCE). Naturally enough, the same value of the G_SOURCE attribute should be shared by all word forms merged into one lexeme as explained in section 2.2. This fact should be controlled during conversion. If it is confirmed, the corresponding information should be expressed in LMF on the level of a ⟨LexicalEntry⟩ instead of a ⟨Word Form⟩ via the originalSource attribute. Note that ISOcat offers standard constants for indicating morphological tools used in the creation of a resource, e.g. derivationTool. Their list is not exhaustive however since no inflectionTool attribute (which would be appropriate for G_INFLECTION_SOURCE) is defined. Hopefully, that situation will evolve with new versions of LMF[6]. In the meantime we use the originalSource twice in order to express both the G_SOURCE and the G_INFLECTION_SOURCE attributes. For instance, the metadata in example (2) could be expressed by the XML feature structure instantiation in Fig. 3.

```
<LexicalEntry>
   <feat att="partOfSpeech" val="subst"/>
   <feat att="originalSource" val="Wikipedia"/>
   <feat att="originalSource" val="Morfeusz"/>
   <Lemma>
      <feat att="writtenForm" val="Bug">
   </Lemma>
   <WordForm>...</WordForm>
</LexicalEntry>
```

Figure 3: Expressing the part of speech and the metadata of the lexeme *Bug*.

The value for the originalSource can be an unresticted string. I suggest to use values *KSNG, Wikipedia, Wikisłownik, World Gazetteer* and *Nazwiska - Futrega* for the first 5 values in section 1.3, and *unknown* for the three values describing the old versions of PNEG.

2.5.3 Semantic Data

The sematic data concerning a ⟨LexicalEntry⟩ are grouped in LMF within the ⟨Sense⟩ element.

Types (GTYPE) and subtypes (G_SUBTYPE) of the PNEG entries stem from the particular projects PNEG was designed for, thus they are obviously not subject to standardization by LMF. LMF allows however to express such data by ⟨MonolingualExternalRef⟩erences which assume the existance of external descriptions (e.g. typologies). For instance, the type and subtype in example (18) could be expressed by the XML instantiation in Fig. 4.

All other attributes in section 1.1 express relations between different entries, however these relations are partly implicit due to the fact that the gazetteer is represented as a flat textual list with no entry identifiers. Thus, example (7) is marked as derived from *Polska*. Determining the precise entry corresponding to such "derivation base" might not always be obvious in a general case. However, PNEG in its present version contains only relative adjectives and inhabitant names stemming from country names. Thus, expressing the corresponding relation in LMF comes down to finding an entry with attribute GTYPE:gaz_country whose lemma is identical to the G_DERIVED_FROM attribute of the current entry.

[6] According to a personal communication with G. Francopoulo

```
<LexicalEntry>
   <Lemma>
      <feat att="writtenForm" val="Najwyższa Izba Kontroli"/>
   </Lemma>
   <Sense>
      <MonolingualExternalRef>
         <feat att="externalSystem" val="PNEG type list"/>
         <feat att="externalReference" val="institution"/>
      </MonolingualExternalRef>
      <MonolingualExternalRef>
         <feat att="externalSystem" val="PNEG subtype list"/>
         <feat att="externalReference" val="national"/>
      </MonolingualExternalRef>
   </Sense>
</LexicalEntry>
```

Figure 4: Referring in LMF to an external typology

Mechanisms supporting semantic relations between different lexical entries are provided in LMF by its **semantics extension**. According to its specifications, the ⟨`Sense`⟩ of a ⟨`LexicalEntry`⟩ contains none or more ⟨`SenseRelation`⟩s with `/targets/` referred to by their sense identifiers. The type of the given relation is described by the attribute `label` with a value being an unrestricted string, which means that the names of relations are not standardized yet. That issue should evolve in future versions of the LMF[7]. Thus, examples (7)–(9) could be expressed by the XML instantiation in Fig. 5.

Similar principles are used for expressing relations represented by attributes G_COUNTRY, G_CITY and G_FULL_FORM in PNEG. For instance entry (5) can be expressed by the XML instantiation in Fig. 6. Note that, for the sake of avoiding redundancy, the attribute G_COUNTRY corresponding to the entry *NIK* (5) has not been explicitly expressed here, since it is represented within its full form lexeme *Najwyższa Izba Kontroli* and can be retrieved for *NIK* via the acronymy relation.

The appendix A shows an extended example corresponding to entries (5)–(9) and (18).

3 Concluding remarks

PNEG is an extensional lexical resource in which most relations between entries are implicit. Restoring these relations makes one face some problems stemming from rendundacies and ambiguities. Thus, it would be methodologically more sound to create “true” databases which would be lexeme-oriented instead of word form-oriented and which would contain explicit relations between lexemes. In such databases, an export towards an LMF format would be more straightforward.

[7] According to a personal communication with G. Francopoulo.

References

[1] M. Becker, W. Drozdzynski, H. Krieger, J. Piskorski, U. Schafer, and F. Xu. Sprout – shallow processing with typed feature structures and unification. In *Proceedings of the International Conference on NLP (ICON 2002)*, Mumbai, India, 2002.

[2] Agnieszka Patejuk and Adam Przepiórkowski. ISOcat definition of the National Corpus of Polish tagset. In *Proceeding of LREC'10 Workshop on LRT Standards, Valletta, Malta*, 2010.

[3] Jakub Piskorski. Named-Entity Recognition for Polish with SProUT. In *LNCS Vol 3490: Proceedings of IMTCI 2004, Warsaw, Poland*, 2005.

[4] Agata Savary and Jakub Piskorski. Language resources for named entity annotation in the national corpus of polish. *Control and Cybernetics*, 40(2), 2010.

[5] Agata Savary and Jakub Piskorski. Lexicons and Grammars for Named Entity Annotation in the National Corpus of Polish. In *Proceeding of IIS'10, Siedlce, Poland*, 2010.

[6] Agata Savary, Joanna Rabiega-Wiśniewska, and Marcin Woliński. Inflection of Polish Multi-Word Proper Names with Morfeusz and Multiflex. *Lecture Notes in Artificial Intelligence*, 5070, 2009.

A An enlarged example of an LMF format for several entries from PNEG

```
<?xml version=’1.0’ encoding="UTF-8"?>
<LexicalResource dtdVersion="15">
  <GlobalInformation>
    <feat att="languageCoding" val="ISO 639-3"/>
    <feat att="characterEncoding" val="UTF-8"/>
  </GlobalInformation>
  <Lexicon>
    <feat att="languageID" val="pol"/>

    <LexicalEntry>
      <feat att="partOfSpeech" val="subst"/>
      <feat att="originalSource" val="KSNG"/>
      <feat att="originalSource" val="Morfeusz"/> <!-- Should be att="inflectionTool" if ISOcat gets extended -->
      <Lemma>
         <feat att="writtenForm" val="Polska"/>
      </Lemma>
      <WordForm>
         <feat att="writtenForm" val="Polska"/>
         <feat att="number" val="sg"/>
         <feat att="gender" val="f"/>
         <feat att="case" val="nom"/>
      </WordForm>
      <WordForm>
         <feat att="writtenForm" val="Polski"/>
         <feat att="number" val="sg"/>
         <feat att="gender" val="f"/>
         <feat att="case" val="gen"/>
      </WordForm>
      <!-- Other word forms -->
      <Sense id="Polska-1">
         <MonolingualExternalRef>
            <feat att="externalSystem" val="PNEG type list"/>
```

```
        <feat att="externalReference" val="country"/>
     </MonolingualExternalRef>
  </Sense>
</LexicalEntry>

<LexicalEntry>
  <feat att="partOfSpeech" val="subst"/>
  <feat att="originalSource" val="unknown"/>
  <Lemma>
     <feat att="writtenForm" val="Najwyższa Izba Kontroli"/>
  </Lemma>
  <WordForm>
     <feat att="writtenForm" val="Najwyższa Izba Kontroli"/>
  </WordForm>
  <Sense id="NajwyzszaIzbaKontroli-1">
     <MonolingualExternalRef>
        <feat att="externalSystem" val="PNEG type list"/>
        <feat att="externalReference" val="institution"/>
     </MonolingualExternalRef>
     <MonolingualExternalRef>
        <feat att="externalSystem" val="PNEG subtype list"/>
        <feat att="externalReference" val="national"/>
     </MonolingualExternalRef>
     <SenseRelation targets="Polska-1">
        <feat att="label" val="meronym"/>
     </SenseRelation>
  </Sense>
</LexicalEntry>

<LexicalEntry>
  <feat att="partOfSpeech" val="subst"/>
  <feat att="originalSource" val="unknown"/>
  <Lemma>
     <feat att="writtenForm" val="NIK"/>
  </Lemma>
  <WordForm>
     <feat att="writtenForm" val="NIK"/>
  </WordForm>
  <Sense id="NIK-1">
     <SenseRelation targets="NajwyzszaIzbaKontroli-1">
        <feat att="label" val="acronym"/>
     </SenseRelation>
  </Sense>
</LexicalEntry>

<LexicalEntry>
  <feat att="partOfSpeech" val="adj"/>
  <feat att="originalSource" val="Wikisłownik"/>
  <feat att="originalSource" val="Morfeusz"/> <!-- Should be att="inflectionTool" if ISOcat gets extended -->
  <Lemma>
     <feat att="writtenForm" val="polski"/>
  </Lemma>
  <WordForm>
     <feat att="writtenForm" val="polski"/>
     <feat att="number" val="sg"/>
     <feat att="gender" val="m3"/>
     <feat att="case" val="nom"/>
  </WordForm>
  <WordForm>
     <feat att="writtenForm" val="polskiego"/>
     <feat att="number" val="sg"/>
     <feat att="gender" val="m3"/>
     <feat att="case" val="gen"/>
  </WordForm>
  <!-- Other word forms -->
  <WordForm>
     <feat att="writtenForm" val="polskie"/>
     <feat att="number" val="sg"/>
     <feat att="gender" val="n"/>
     <feat att="case" val="voc"/>
  </WordForm>
  <!-- Other word forms -->
  <Sense id="polski-1">
    <SenseRelation targets="Polska-1">
```

```
           <feat att="label" val="relativeAdjective"/>
        </SenseRelation>
      </Sense>
    </LexicalEntry>

    <LexicalEntry>
      <feat att="partOfSpeech" val="subst"/>
      <feat att="originalSource" val="Wikisłownik"/>
      <feat att="originalSource" val="Morfeusz"/>
      <Lemma>
        <feat att="writtenForm" val="Polak"/>
      </Lemma>
      <WordForm>
        <feat att="writtenForm" val="Polak"/>
        <feat att="number" val="sg"/>
        <feat att="gender" val="m1"/>
        <feat att="case" val="nom"/>
      </WordForm>
      <WordForm>
        <feat att="writtenForm" val="Polakowi"/>
        <feat att="number" val="sg"/>
        <feat att="gender" val="m1"/>
        <feat att="case" val="dat"/>
      </WordForm>
      <WordForm>
        <feat att="writtenForm" val="Polacy"/>
        <feat att="number" val="pl"/>
        <feat att="gender" val="m1"/>
        <feat att="case" val="nom"/>
      </WordForm>
      <!-- Other word forms -->
      <Sense id="Polak-1">
        <SenseRelation targets="Polska-1">
          <feat att="label" val="relativePersonName"/>
        </SenseRelation>
      </Sense>
    </LexicalEntry>

    <LexicalEntry>
      <feat att="partOfSpeech" val="depr"/>
      <feat att="originalSource" val="Wikisłownik"/>
      <feat att="originalSource" val="Morfeusz"/>
      <Lemma>
        <feat att="writtenForm" val="Polak"/>
      </Lemma>
      <WordForm>
        <feat att="writtenForm" val="Polaki"/>
        <feat att="number" val="pl"/>
        <feat att="gender" val="m2"/>
        <feat att="case" val="nom"/>
      </WordForm>
      <WordForm>
        <feat att="writtenForm" val="Polaki"/>
        <feat att="number" val="pl"/>
        <feat att="gender" val="m2"/>
        <feat att="case" val="voc"/>
      </WordForm>
      <!-- Other word forms -->
      <Sense id="Polak-2">
        <SenseRelation targets="Polska-1">
          <feat att="label" val="relativePersonName"/>
        </SenseRelation>
        <SenseRelation targets="Polak-1">
          <feat att="label" val="depreciativeVariant"/>
        </SenseRelation>
      </Sense>
    </LexicalEntry>

  </Lexicon>
</LexicalResource>
```

```
<LexicalEntry>
   <feat att="partOfSpeech" val="adj"/>
   <Lemma>
      <feat att="writtenForm" val="polski"/>
   </Lemma>
   <Sense id="polski1">
      <SenseRelation targets="Polska-1">
         <feat att="label" val="relativeAdjective"/>
      </SenseRelation>
   </Sense>
</LexicalEntry>
<LexicalEntry>
   <feat att="partOfSpeech" val="subst"/>
   <Lemma>
      <feat att="writtenForm" val="Polak"/>
   </Lemma>
   <Sense id="Polak-1">
      <SenseRelation targets="Polska-1">
         <feat att="label" val="relativePersonName"/>
      </SenseRelation>
   </Sense>
</LexicalEntry>
<LexicalEntry>
   <feat att="partOfSpeech" val="depr"/>
   <Lemma>
      <feat att="writtenForm" val="Polak"/>
   </Lemma>
   <Sense id="Polak-2">
      <SenseRelation targets="Polska-1">
         <feat att="label" val="relativePersonName"/>
      </SenseRelation>
      <SenseRelation targets="Polak-1">
         <feat att="label" val="depreciativeVariant"/>
      </SenseRelation>
   </Sense>
</LexicalEntry>
<LexicalEntry>
   <Lemma>
      <feat att="writtenForm" val="Polska"/>
   </Lemma>
   <Sense id="Polska-1">
      <MonolingualExternalRef>
         <feat att="externalSystem" val="PNEG type list"/>
         <feat att="externalReference" val="country"/>
      </MonolingualExternalRef>
   </Sense>
</LexicalEntry>
```

Figure 5: Expressing semantic relations between derivatives *polski* and *Polak* (possibly depreciative) and their derivation base *Polska*.

```
<LexicalEntry>
   <Lemma>
      <feat att="writtenForm" val="NIK"/>
   </Lemma>
   <Sense id="NIK1">
      <SenseRelation targets="NajwyzszaIzbaKontroli1">
         <feat att="label" val="acronym"/>
      </SenseRelation>
   </Sense>
</LexicalEntry>
<LexicalEntry>
   <Lemma>
      <feat att="writtenForm" val="Najwyższa Izba Kontroli"/>
   </Lemma>
   <Sense id="NajwyzszaIzbaKontroli1">
      <SenseRelation targets="Polska-1">
         <feat att="label" val="meronym"/>
      </SenseRelation>
   </Sense>
</LexicalEntry>
<LexicalEntry>
   <Lemma>
      <feat att="writtenForm" val="Polska"/>
   </Lemma>
   <Sense id="Polska-1">
      <MonolingualExternalRef>
         <feat att="externalSystem" val="PNEG type list"/>
         <feat att="externalReference" val="country"/>
      </MonolingualExternalRef>
   </Sense>
</LexicalEntry>
```

Figure 6: Expressing a semantic relation between an institution name acrynom (*NIK*), its full form (*Najwyższa Izba Kontroli*) and its holonym country (*Polska*).