# Enhancing an English-Polish electronic dictionary for multiword expression research

**Piotr Bański**
Institute of English Studies, University of Warsaw
Nowy Świat 4, 00-497 Warszawa, Poland
E-mail: bansp@o2.pl

**Radosław Moszczyński**
Department of Formal Linguistics, University of Warsaw
Browarna 8/10, 00-311, Warszawa, Poland
E-mail: r.moszczynski@uw.edu.pl

**Abstract**

This paper describes a project aimed at converting a legacy representation of English idioms into an XML-based format. The project is set in the context of a large electronic English-Polish dictionary which contains several hundred formalized idiom descriptions and which has been released under the terms of a free license. In short, the project consists of three phases: cleaning up the dictionary markup, extracting the legacy idiom representations, and converting them into TEI P5 XML constrained by a RelaxNG grammar created for this purpose and constituting a module that can be included as part of the TEI P5 schema. The paper contains general descriptions of the individual phases and several examples of XML-encoded idioms. It also suggests some directions for further research, which include abstracting the XML-ized idiom representations into general syntactic patterns and using the representations to automatically identify idioms in tagged corpora.

## 1. Introduction

The project described here targets an English-Polish dictionary compiled by Tadeusz Piotrowski and Zygmunt Saloni, recently made available to the public in the electronic form. Due to its free license and rich linguistic information that includes a formal description of English idioms, the dictionary constitutes a valuable resource for both dictionary users and linguists. Our project has two main aims: (i) fixing various markup errors that the dictionary contains after two conversions from its original TeX form — the result will be a fully standardized and interchangeable XML format conformant with the TEI P5 Guidelines (Sperberg-McQueen & Burnard, 2007), as well as (ii) extending the description of idioms — fleshing out their structural properties and at the same time making it possible to describe the individual constituents, with a view to facilitating further research in this domain.

The dictionary was originally created as a traditional publication that appeared in print as Piotrowski and Saloni (1992). Its initial electronic format was pure TeX. In the late 1990s, it was converted into SGML for the purposes of the STEEL (Developing Specialized Translation/Foreign Language Understanding Tools for Eastern European Languages) project, which involved technologies developed by Xerox — most notably IDAREX (IDioms As REgular eXpressions). The exact SGML format is described in detail in Głowińska & Woliński (2000). Piotrowski (1999) provides background information on the conversion process and some design decisions behind the original version of the dictionary.

Recently, the dictionary has been released by the authors as TeX, SGML and XML source, dual-licensed under GNU GPL 2 and GNU FDL 1.2. Around the same time an XML version of the dictionary has been created as a term assignment for a course held at the Institute of Informatics, University of Warsaw. The target markup scheme was early TEI P5.

The XML version of the dictionary contains 16002 main entries with 1529 subentries. Its distinguishing feature is the inclusion of IDAREX formulas describing multiword expressions (broadly, idioms and strong collocations) that the given lexeme can be part of. The XML version is not flawless, but the conversion was successful enough to constitute the basis for further work, which consists of three main phases, listed immediately below and described in more detail in the remainder of the paper:

- Cleaning up the existing markup and preparing a TEI P5 customization for it, in the form of a RELAX NG schema, aiming at full P5 conformance within the TEI namespace; the deliverable of this stage will be the dictionary itself, as well as the customization (so-called ODD) file, from which the schema is derived. This phase is nearly completed, pending addition of entries lost in the TeX-to-SGML conversion (see below).
- Recoding the multiword-expression formulas as chunks of XML — we have a small RELAX NG regular expression grammar designed to encode the existing IDAREX formulas and extend them beyond their current limits (see section 3).
- Establishing methods of visualizing the formulas within dictionary entries together with methods allowing for enhanced navigation, e.g. across structures of the same syntactic type.

The resource created this way will be used as a modern electronic dictionary, but at the same time it will also serve as a basis for NLP applications (e.g. idiom recognition in corpora).

## 2. Clean-up phase

During the clean-up phase, we focused on three types of errors that appeared in the dictionary as a result of the conversion processes that it has undergone. Firstly, we had to handle semantic errors, i.e. violations of the semantic constraints of the TEI abstract model (see Sperberg-McQueen & Burnard, 2007: ch. 23) resulting

from inappropriate usage of elements from the TEI namespace. For example, <xr> elements contained subentries or expansions of abbreviations instead of providing cross-references.

The second type of errors that needed to be fixed were logical errors — for example, 624 entries (including numerous abbreviations but also words such as e.g. *asleep*, *behead*, *mall*, *zodiac*) lack part-of-speech (POS) labels. This was an error carried over from the SGML version — during the TeX-to-SGML conversion, POS labels had been added automatically to most, but not all, entries/senses. In the cleaned-up version, POS labels have been added to entries other than those describing abbreviations.

Finally, plain conversion errors still need to be handled. On the way from TeX to SGML, some entries had been omitted, among them *a*, *can*, *die*, *keep*, *put*, *take*, *under*, i.e. words from the top regions of any frequency list (altogether, 761 entries have not made it into the XML version). This last issue is what hinders the completion of the clean-up phase, as it is not merely the matter of recoding the missing entries in XML but also of recovering the IDAREX specifications for the idioms that they are part of — these specifications are not available to us yet. The legacy IDAREX representations themselves also require corrections — we return to this issue below

## 3. IDAREX formalism

For the purposes of the STEEL project, all idioms in the original dictionary by Piotrowski and Saloni had to be encoded as IDAREX formulas. Most of the work was done automatically, with some residue left for human lexicographers. In what follows, we look at three English multiword expressions encoded with IDAREX (the formalism is described in detail in Breidt, Segond & Valetto, 1996). The first formula describes exactly two phrases: *at the double* and *on the double*.

```
1   [:at | :on] :the :double
```

In short, a prepended colon marks a surface form (on the "output tape", in the finite-state transducer terminology), a postpended colon marks what in the Two-Level Morphology framework of Kimmo Koskenniemi (see Karttunen & Beesley, 2001) is called the lexical level ("input tape"). In general, surface level tokens are unchangeable, while words at the lexical level can be inflected. The vertical bar stands for alternative and square brackets are used for grouping tokens.

Example (1) can be expressed informally (with omitted namespace prefixes and suppressed attributes) in our XML notation as follows:

```
2
<sequence>
    <choice>
        <surface>at</surface>
        <surface>on</surface>
    </choice>
    <surface>the</surface>
    <surface>double</surface>
</sequence>
```

In the next expression, "ADV*" signals that an adverb may be used any number of times, and "take Verb:" is used to restrict the word *take* to its verbal sense and to signal that it can appear in any of its inflectional forms:

```
3
ADV* take Verb: :the :bull :by :the :horns
```

The expression can be recoded into XML as follows:

```
4
<sequence>
    <zeroOrMore>
        <var>ADV</var>
    </zeroOrMore>
    <lexical pos="V">take</lexical>
    <surface>the</surface>
    <surface>bull</surface>
    <surface>by</surface>
    <surface>the</surface>
    <surface>horns</surface>
</sequence>
```

Note that the token "take Verb:" is now recoded as a single XML element with an appropriate @pos attribute. Note also that in our formalism, the @pos attributes of variables and lexical tokens are uniform, which is a basis for further applications (e.g. the establishment of possible references and substitutions). Because we want to proceed step-by-step, the @pos attributes for <surface> elements are for the time being optional. The same is true of @entry attributes which (after lemmatization) will point at the relevant entries of the dictionary.

Our final example is a formula requiring a choice of one out of two related sequences. This time the XML in (6) overleaf is pasted from an actual fragment of the dictionary, where it is located in the <gramGrp> element, storing grammatical information. The dictionary with such extensions is valid against a version of TEI P5 1.0 Guidelines with our additional IDAREX module.

An additional virtue of this representation is the possibility of relating phrasal and word-level categories by means of the @pos and @extent attributes. Something that we want to implement soon is a possibility of relating elements of such alternative sequences (notice that the two NP symbols in (5) designate the same object). The XPointer element() function seems an excellent tool for this purpose.

```
5   [take V: NP :into :account|
    take V: :account :of NP]
```

6

```
<gramGrp>
    <idx:idarex>
        <idx:choice>
            <idx:sequence>
    <idx:lexical pos="v">take
    </idx:lexical>
    <idx:var pos="n"
    extent="phrase">NP</idx:var>
    <idx:surface>into</idx:surface>
    <idx:surface>account</idx:surface>
            </idx:sequence>
            <idx:sequence>
<idx:lexical pos="v">take</idx:lexical>
<idx:surface>account</idx:surface>
<idx:surface>of</idx:surface>
<idx:var pos="n"
extent="phrase">NP</idx:var>
            </idx:sequence>
        </idx:choice>
    </idx:idarex>
</gramGrp>
```

As we have mentioned, the IDAREX formulas used in the dictionary require a cleanup before we can start automatic conversion from strings to XML: some variable labels appear to be ad hoc (e.g. the labels MDWD and MDCL, encoding some aspects of modal verbs, are not defined and used only once in the entire dictionary). The IDAREX formulas are also not entirely uniform and they do not always reflect the structure of idioms. We omit the examples for lack of space, mentioning only that e.g. the abbreviation "sth" is symbolized by the variables NTH, NP or N, or is sometimes ignored altogether.

At this stage of the project, in our test dictionary, we keep the XML-ized IDAREX specifications inside entries, as content of the appropriate `re/gramGrp/gram` elements, in order to keep them where they belong, according to their lexical and semantic relationships. At a later stage of the project, we will consider externalizing the IDAREX level of annotation to a separate file, as schematically illustrated below.

7

dictionary.xml (TEI P5, with cross-references pointing at chunks of idarex.xml)

↕

idarex.xml (chunks of structures that elements of the dictionary refer to, possibly in a many-to-one fashion)

This kind of representation will be helpful in identifying common syntactic patterns, both for the purpose of subcategorizing idioms according to the structures they appear in, and in order to facilitate further research (we return to this issue in section 4). Also with a view towards further research, the IDAREX statements will be categorized according to Moszczyński's (2007) draft typology of multiword expressions, which requires empirical testing on a large set of data, such as the idioms in Piotrowski and Saloni's dictionary. As can be seen in examples 2 and 4, our first move is to encode the tokens of XML-ized IDAREX formulas at the same level of abstraction as in the IDAREX strings. Later on, each token element will receive additional attributes specifying its POS and pointing at the dictionary entry where it is described.

The final stage of the project will involve the cosmetic issue of visualizing the multiword formulas in a user-friendly way, possibly with a reference from each word to the entry where this word is described. At this phase, minor housekeeping may also be performed: for example, in the TeX source, the idiom *keep tabs on sb* is mentioned under the headword *tab* but not under *keep* — in the final version, we want it to appear in both places.

## 4. Impact of the project and further research

The dictionary, even without the IDAREX formulas, is the largest free electronic English-Polish dictionary prepared by expert lexicographers that has been made available to the public so far, and we consider the initial cleanup stage of our project as partial payment of the debt that the community owes to Tadeusz Piotrowski and Zygmunt Saloni for releasing their work under a free license. By making the dictionary fully TEI-conformant and adding the missing entries, we hope to increase its value. The cleaned-up dictionary will become part of the FreeDict project (http://freedict.org/), which hosts bilingual electronic dictionaries.

There are several advantages of recoding IDAREX formulas in XML:

- Representation: instead of strings that for each parse would have to be tokenized in a non-trivial way (cf. example 3, where "take Verb:" is a single token), the XML representation makes it possible to encode some structural aspects of the IDAREX (by means of the `<choice>` or `<zeroOrMore>` elements).
- Atomization: each element of the XML representation can be furnished with the part of speech that it represents and with a hyperlink to the entry/sense where it is described, as mentioned in section 3.
- The next logical step in the research is extracting common patterns from XML representations (at this stage, the IDAREX descriptions may be moved to a separate file, cross-linked from the dictionary, cf. 7); this means abstracting from concrete tokens toward part-of-speech variables.
- Armed with this kind of representations, one may attempt automatic identification of idioms in corpora.
- The representation may also be generalized to cover all valence relationships, for the purpose of

creating valence codes for a syntactic description of lexemes.

We are planning to release the cleaned-up English-Polish dictionary for FreeDict and to release a version with recoded IDAREX formulas later this year.

## 5. References


Breidt, Elisabeth; Segond, Frédérique; Valetto, Giuseppe. (1996). Formal Description of Multi-Word Lexemes with the Finite-State Formalism IDAREX. In *Proceedings of the 16th Conference on Computational Linguistics*, Volume 2.Morristown, NJ: Association for Computational Linguistics, pp. 1036–1040. http://acl.ldc.upenn.edu/C/C96/C96-2182.pdf.

Głowińska, Katarzyna & Woliński, Marcin. (2000). Angielsko-polski słownik elektroniczny XeLDA. Acta Universitatis Nicolai Copernici. Studia Slavica 5, no. 343, pp. 119–124.

Karttunen, Lauri & Beesley, Kenneth R.. (2001). A Short History of Two-Level Morphology. ESSLLI-2001 Special Event. Available at http://www2.parc.com/istl/members/karttune/publications/esslli-2001/twol-history.pdf

Moszczyński, Radosław. (2007). A practical classification of multiword expressions. In *Proceedings of the ACL 2007 Student Research Workshop*. Available at http://acl.ldc.upenn.edu/P/P07/P07-3004.pdf

Piotrowski, Tadeusz. (1999). Tagging and Conversion of a Bilingual Dictionary for XeLDA, a Xerox Computer Assisted Translation System. In *Papers in Computational Lexicography COMPLEX '99 Proceedings*. Budapest: Hungarian Academy of Sciences, pp. 113-120.

Piotrowski, Tadeusz & Saloni, Zygmunt. (1992). *Nowy słownik angielsko-polski polsko-angielski [New English-Polish, Polish-English dictionary]*. Warszawa: Editions Spotkania.

Sperberg-McQueen, C.M. & Burnard, Lou (eds). (2007). TEI P5: Guidelines for Electronic Text Encoding and Interchange. Available at http://www.tei-c.org/P5/