Quality Control for Wordnet Development

Pavel Smrž

Faculty of Informatics, Masaryk University in Brno
Botanická 68a, 602 00 Brno, Czech Republic
Email: smrz@fi.muni.cz

Abstract. This paper deals with quality assurance procedures for general-purpose language resources. Special attention is paid to quality control in wordnet development. General issues of quality management are tackled; technical as well as methodological aspects are discussed. As a case study, the application of the described procedures is demonstrated on the quality evaluation techniques in the context of the BalkaNet project.

1 Introduction

The BalkaNet project [1] aims at the development of wordnet-like lexical semantic networks for Czech and 5 Balkan languages – Bulgarian, Greek, Romanian, Serbian, and Turkish. As it shares many fundamental principles with the EuroWordNet project [2], it has been expected to employ the same procedures, policy, structure and tools as the previous project. However, discovered limitations of the EuroWordNet approach brought us to the decision to change data format, to design and implement new applications, and also to propose a modified perspective of the future development of the lexical semantic databases. Our conception, structure and tools are currently applied not only by members of the BalkaNet consortium but also by many other teams developing lexical databases all over the world.

There are many application-specific language resources developed with the goal to be directly integrated in a particular environment. On the other hand, there are resources that have been used or aim at their application in various NLP tasks. WordNet is the most prominent example. Though created to model human mental lexicon it has been employed in many domains from information retrieval to cultural linguistics, from text classification to language teaching, word-sense disambiguation, machine translation, etc.

Many well-established methods are available to evaluate the quality and contribution of language resources for specific application tasks. For example, the standard precision/recall graphs or F-measures are the most popular in the information retrieval. The fields of evaluation machine translation or information extraction systems pay also traditionally a strong attention to the quality assurance.

The procedures of quality control for general-purpose language resources are much less known. Moreover, the results of our research clearly show that this area has been strongly underestimated in many previous projects. Another finding suggests that if quality assurance policy has not been applied the results could differ considerably from that what was declared.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 206–212.

2 General Considerations

The most obvious requirement for a resource that aims at general usage is the availability of documentation of the process of its development and the final state of data. Resource documentation should be comprehensive but at the same time concise to allow quick scan. Unfortunately, many language resources resulting from various research projects account the role to a set of the standard project deliverables. In addition to the fact that these documents are often longer than necessary and do not describe all aspects of the resource, this approach does not reflect the process of development. Deliverables correspond to the state of knowledge and development of the resource at a particular time. Decisions and views can change during the project. The best strategy is therefore to summarize the description of resources in the end of such projects and check validity of information in all documents that will be part of the documentation.

The terminology used in the resource description should be also explicitly defined. Even the meaning of terms that seem to be basic in the context should be tackled. For example, synonymic set – synset – is the fundamental building block of wordnets but still it should be precisely described what kinds of variants (typographic, regional, register ...) will be contained in a synset. The Princeton WordNet itself is not entirely consistent in this respect – lake, loch and lough – as regional variants of the same concept – form 3 different synsets, lake is the hypernym of the two others.

The description of the data format in which the resource is provided plays also a crucial role. As XML has become de facto standard for data interchange, it is natural to make data available in XML and release the relevant DTD description. Data types of XML entities and other constraints on the tag content should be also specified. Elaborate standards from "the XML family", e. g. XML Schema [3] can be used to formally capture these definitions.

Along with the description of the data format it is appropriate to publish quantitative characteristics of the created data. A special attention should be paid to empty tags in the case of XML representation as it may signalize data inconsistency.

Our experience in previous projects aiming at development of language resources clearly showed that one of the most successful procedures to control the quality of linguistic output is to implement a set of validation checks and regularly publish their results. It holds especially for projects with many participants that are not under the same supervision. Validation check reports together with the quantitative assessment can serve as development synchronization points too.

3 Case Study of Quality Control in BalkaNet

The BalkaNet project will run till August 2004. Thus, we are not able to present the final documentation of all decisions that have been made in the course of the multilingual wordnet development. However, we present the current state of the project which reflects the refined quality control policy the BalkaNet consortium has adopted.

All partners agreed to prepare and update "resource description sheet" for the wordnet they develop. Such a specification should contain at least:

- description of the content of synset records and constraints on data types;

- types of relations included together with examples;
- degree of checking relations borrowed from PWN (see the note about the expand model below);
- numbering scheme of different senses (random, according to their frequency in a balanced corpus, from a particular dictionary, etc.)
- source of definitions and usage examples;
- order of literals in synsets (corpus frequency, familiarity, register or style characteristics).

One of the main characteristics that holds from very beginning of BalkaNet is the focus on large-scale overlap between national wordnets. The goal of this approach is to maximize the possibility of future applicability of the created database as a whole. A special set of synsets – BCS (BalkaNet Common Synsets) has been chosen and all partners agreed on the schedule of the gradual development. Several criteria have been adopted in the BCS selection process, which has taken the following steps:

1. All synsets contained in EuroWordNet base concepts have been included to maximize the overlap between the two projects.
2. The set has been extended based on the proposals of all partners who added synsets corresponding to the most frequent words in corpora and in various dictionary definitions for their particular languages.
3. As an additional criterion, several noun synsets that had many semantic relations in the Princeton WordNet database have been added.
4. All the selected synsets based on PWN 1.5 have been automatically mapped to PWN 1.7.1, which is currently the version BalkaNet is connected to. The synsetss that found one-to-one correspondence in the new version have been finally chosen.
5. All the hypernyms and holonyms of the chosen synsets have been added to BCS as it was decided to close the set in this respect.

All the steps (except the second for the proposer) imply the adoption of expand model for building a substantial part of the national wordnets. However, there is still room for the merge model, e. g. a significant portion of verb synsets in the Czech wordnet originated that way.

Synsets are formed by true context synonyms as well as variants (typographic, regional, style, register ...) in the BalkaNet wordnets. Moreover, verb synsets contain literals linked by a rich set of relations, e. g. aspect opposition and iteratives.

All the data should be linked to PWN till the end of the project. BalkaNet started with the idea to provide correspondence with PWN 1.5 and thus be compatible with EuroWordNet. However, the discovered limitations of PWN 1.5 let to the switch to PWN 1.7.1 which is much more consistent. As the new PWN 2.0 has been released in the last months the possibility of automatic re-linking of BalkaNet data to this version will be investigated too.

All national wordnets share the same data structure in XML. A synset described in this notation could look like:

```
<SYNSET>
  <ID>ENG171-08299742-n</ID> <POS>n</POS>
  <SYNONYM>
    <LITERAL>front man<SENSE>1</SENSE></LITERAL>
```

```
    <LITERAL>front<SENSE>8</SENSE></LITERAL>
    <LITERAL>figurehead<SENSE>1</SENSE></LITERAL>
    <LITERAL>nominal head<SENSE>1</SENSE></LITERAL>
    <LITERAL>straw man<SENSE>1</SENSE></LITERAL>
  </SYNONYM>
<ILR><TYPE>hypernym</TYPE>ENG171-08207586-n</ILR>
<DEF>a person used as a cover for some questionable activity</DEF>
</SYNSET>
```

The corresponding DTD for all BalkaNet wordnets then looks like:

```
<!ELEMENT WORDNET - - (SYNSET*) >
<!ELEMENT SYNSET  - - (ID, POS, SYNONYM, ILR*, ELR*, BCS?,
                       DEF?, USAGE*, SNOTE*, STAMP?) >

<!ELEMENT SYNONYM - - (LITERAL+) >
<!ELEMENT LITERAL - - (#PCDATA, SENSE, LNOTE?) >
<!ELEMENT SENSE   - - (#PCDATA) >
<!ELEMENT LNOTE   - - (#PCDATA) >

<!ELEMENT ILR  - - (TYPE, #PCDATA) >
<!ELEMENT ELR  - - (TYPE, #PCDATA) >
<!ELEMENT TYPE - - (#PCDATA) >

<!ELEMENT ID    - - (#PCDATA) >
<!ELEMENT POS   - - (#PCDATA) >
<!ELEMENT BCS   - - (#PCDATA) >
<!ELEMENT DEF   - - (#PCDATA) >
<!ELEMENT USAGE - - (#PCDATA) >
<!ELEMENT SNOTE - - (#PCDATA) >
<!ELEMENT STAMP - - (#PCDATA) >
```

The ID tag acts as the primary key of the entries and is also used in links where it substitutes the verbosity of proper XML linking mechanisms [4,5,6]. Identifiers are found in two slightly different forms:

1. Synsets connected to PWN are identified by three-part strings – the first is the version identifier (e. g. ENG15 for PWN version 1.5), the second is the offset in the PWN files for nouns, adjectives, verbs, or adverbs, and the third one is the concrete POS.
2. Synsets added by the consortium partners start with the three-letter language identifiers that correspond to the international standard ISO 639-2. The following number is generated sequentially to ensure uniqueness.

The second mentioned group is just a matter of the progressive development of national wordnets. Most of the synsets will be linked to their English equivalents till the end of the project. It means they will get IDs from PWN. The rest will form the core of what is called BalkaNet ILI (Inter-Language Index), or BILI. The prefix will be BWN10 and English

definition will be provided. The most discussed examples of this type so far are the names of meals served in the Balkan region.

A special mechanism has been adopted to signalize lexical gaps – concepts that are not lexicalized in a language. Such entries are labeled <NL/> in the BalkaNet database and they should be ignored when working with a particular wordnet as a monolingual resource.

The current DTD complies with the needs of the development process (BCS tags for synchronization, STAMP tag for management purposes, etc.). The final version will probably eliminate these tags and maybe adds others to facilitate linking to other resources.

Simple scripts using standard utilities like sort or diff tools have been implemented to compute quantitative characteristics. All the XML files are first normalized to eliminate effects of the different structure. The following frequency values are then computed:

- tag frequencies;
- ratio of the number of literals in the national wordnet and in PWN;
- ID prefix frequencies;
- frequency of link types;
- frequency of POS;
- coverage of BCS;
- number-of-senses distribution;
- number of "multi-parent" synsets;
- number of leaves, inner nodes, roots, free nodes in hyper-hyponymic "trees";
- path-length distribution.

Table 1 captures the most interesting statistics that reflect the state of Balkanet development in the end of the second year of the project.

Table 1. Current statistics on wordnets developed in BalkaNet

Wordnet	Bulgarian	Czech	Greek	Romanian	Serbian	Turkish	Princeton
Synsets	13,425	25,453	13,523	11,698	4,557	9,509	111,223
Literals	24,118	37,883	17,759	23,571	7,891	14,382	195,817
Lit/Syn	1.80	1.49	1.31	2.01	1.73	1.51	1.76
BCS	8,496	7,525	5,427	6,744	4,307	7,391	8,496

4 Automatic and Semi-automatic Quality Checking

The quality control has been one of the priorities of the BalkaNet project. As our evaluation proves even the actual data from the second year of the project are more consistent than the results of previous wordnet-development projects. Part of the success story definitely lies in the implementation of strict quality control and data consistency policy.

Data consistency checks can be considered from various points of view. They can be fully automatic or need less or more manual effort. Even if supported by software tools, manual checks present tedious work that moreover needs qualified experts. Another criterion for

applicability of checks is whether they can be applied to all languages or they are language-specific (e. g. constraints on characters from a particular codepage). An important issue is also the need for additional resources and/or tools (e. g. annotated monolingual or parallel corpora, spell-checkers, explanatory or bilingual dictionaries, encyclopedias, lemmatizers, morphological analyzers).

Similarly to the scripts for quantitative characteristics we have developed a set of checks that validate wordnet data in the XML format. The following inconsistencies are regularly examined on all BalkaNet data:

- empty ID, POS, SYNONYM, SENSE (XML validation);
- XML tag data types for POS, SENSE, TYPE (of relation), characters from a defined character set in DEF and USAGE;
- duplicate IDs;
- duplicate triplets (POS, literal, sense);
- duplicate literals in one synset;
- not corresponding POS in the relevant tag and in the ID postfix;
- hypernym and holonym links (uplinks) to a synset with different POS;
- dangling links (dangling uplinks);
- cycles in uplinks (conflicting with PWN, e. g. "goalpost:1" is a kind of post is a kind of "upright:1; vertical:2" which is a part of "goalpost:1");
- cycles in other relations;
- top-most synset not from the defined set (unique beginners) – missing hypernym or holonym of a synset (see BCS selecting procedure above);
- non-compatible links to the same synset;
- non-continuous numbering where declared (possibility of automatic renumbering).

The results of the checks are also regularly sent to the developers that are responsible for corrections. The current practice will be probably even further simplified when a new tool for consistency checking with a user-friendly graphical interface will be developed.

Semi-automatic checks that need additional language resources to be integrated are usually performed by each partner depending on the availability of the resources:

- spell-checking of literals, definitions, usage examples and notes;
- coverage of the most frequent words from monolingual corpora;
- coverage of translations (bilingual dictionaries, parallel corpora);
- incompatibility with relations extracted from corpora, dictionaries, or encyclopedias.

In addition to the above-mentioned checks, BalkaNet developers often work with outputs of various pre-defined queries retrieving "suspicious" synsets or cases that could indicate mistakes of lexicographers. For examples, these queries can list:

- nonlexicalized literals;
- literals with many senses;
- multi-parent relations;
- autohyponymy, automeronymy and other relations between synsets containing the same literal;
- longest paths in hyper-hyponymic graphs;

- similar definitions;
- incorrect occurrences of defined literals in definitions;
- presence of literals in usage examples;
- dependencies between relations (e. g. near antonyms differing in their hypernyms);
- structural difference from PWN and other wordnets.

Besides all the mentioned validation checks, quality of created resources is evaluated in their application. Several partners already used their data to annotate corpus text for WSD experiments. Such an experience usually shows missing senses or impossibility to choose between different senses. Another type of work that helps us to refine information in our wordnet was the comparison between the semantic classifications from the wordnet with the syntactic patterns based on computational grammar.

5 Conclusions and Future Directions

It is obvious that the effort aiming at the quality of developed resources paid already off in the form of consistent resulting data that can be successfully used in various applications. The BalkaNet project will follow the started approach and the set of consistency checks used to validate wordnets will be published in its end.

We will try to test and generalize the GUI tool for validation checking mentioned above. We will also continue to develop the XML based application that will employ XSLT and other XML standards to define the tests [7].

Acknowledgements

This work was supported by Ministry of Education of the Czech Republic Research Intent CEZ:J07/98:143300003 and by EU IST-2000-29388.

References

1. Balkanet project website, `http://www.ceid.upatras.gr/Balkanet/`.
2. Eurowordnet project website, `http://www.illc.uva.nl/EuroWordNet/`.
3. Fallside, D. C.: XML Schema Part 0: Primer (2001) `http://www.w3.org/TR/xmlschema-0/`.
4. DeRose, S., Maler, E., Orchard, D.: XML Linking Language (XLink) Version 1.0 (2001) `http://www.w3.org/TR/xlink`.
5. DeRose, S., Jr., R. D., Grosso, P., Maler, E., Marsh, J., Walsh, N.: XML Pointer Language (XPointer) W3C Working Draft (2002) `http://www.w3.org/TR/xptr`.
6. Clark, J., DeRose, S.: XML Path Language (XPath) Version 1.0 (1999) `http://www.w3.org/TR/xpath`.
7. Smrž, P., Povolný, M.: Deb – dictionary editing and browsing. In: Proceedings of the EACL03 Workshop on Language Technology and the Semantic Web: The 3rd Workshop on NLP and XML (NLPXML-2003), Budapest, Hungary (2003) 49–55.