

OpenThesaurus: Building a Thesaurus with a Web Community

Daniel Naber <daniel.naber@t-online.de>

2004-06-13

Abstract

OpenThesaurus is a database-driven website to collect synonyms of the German language. People can sign up and search the database for words or add meanings and synonyms. The collected data – currently more than 30,000 words – is freely available under an Open Source license. Ready-to-use thesaurus files for OpenOffice.org can be downloaded. The PHP files that drive the website can also be used to provide the same service for other languages. The website’s URL is <http://www.openthesaurus.de>.

Contents

1	Introduction	1
2	The OpenThesaurus Website	2
2.1	Initial Data Import	2
2.2	Database Structure	2
2.3	User Interface	3
2.4	Administration Area	4
2.5	Data Export	5
2.6	Using OpenThesaurus for Languages other than German	5
2.7	Statistics	5
3	Conclusion	6
A	Document History	6

1 Introduction

A thesaurus is a dictionary that lists words which have a similar or related meaning. The most simple case is a pair of synonyms, i.e. two words which have the same meaning. An example in German is *Samstag* and *Sonnabend* (both meaning *Saturday*). An English example is *car* and *automobile*.

A thesaurus can be useful for several purposes: It can serve as a simple dictionary-replacement that explains the meaning of a word by listing other words with the same or with a similar meaning. It can be used by writers to produce better texts, for example by looking up a more suitable adjective or by looking up a word that is easier to understand as a replacement for a foreign term. Knowledge about synonyms might also be used by search engines to find documents that contain information about the subject one is looking for, but that uses different terminology.

Before OpenThesaurus was established there has been no freely available thesaurus for the German language. The OpenThesaurus project has developed a thesaurus that is available under the terms of the GNU General Public License [GPL]. The structure is similar to WordNet, which is a freely available English lexical database “whose design is inspired by current psycholinguistic theories of human lexical memory” [WordNet]. From the OpenThesaurus point-of-view, the most important feature of WordNet is that its data is organized in so-called synonym sets (synsets), each synonym set representing one meaning. For example, WordNet contains a synonym set *car, auto, automobile, machine, motorcar*. All these words refer to the

same group of objects (“4-wheeled motor vehicle” – WordNet knows that, too), i.e. they are synonyms. Of course, *machine* may also be used with a different meaning. In fact, if you look up *machine* in WordNet, you will get eight different meanings, one of them being the synonym set quoted above.

OpenThesaurus basically works the same way, but for the German language. While WordNet supports many different relations between its synonym sets (antonymy, meronymy etc), OpenThesaurus is, for now, limited to synonym sets and superordinate/subordinate relations between synonym sets. The synonyms of a synonym set have no order, but they are displayed in alphabetical order.

2 The OpenThesaurus Website

2.1 Initial Data Import

The number of words in natural languages is almost unlimited. For example, the German [Duden] contains more than 115,000 words, but it does not even include many proper nouns. A thesaurus needs to contain at least a few thousand entries to be useful. So one should try to avoid starting from scratch, but import existing data instead:

There are two projects, [GermaNet] and [Wortschatz Deutsch], which have collected a large number of German synonyms. However, these projects do not offer their results under an Open Source license, so they cannot be used for OpenThesaurus.

One could try to import the words from the German ispell dictionary, which is freely available. Unfortunately this dictionary does not contain any information about synonyms, so one would have many words but not a single synonym.

We finally generated the data from the German/English dictionary [Dict] which is available under the GPL. The dictionary has about 115,000 entries. Some of them look like this:

```
Bandit {m}; Räuber {m} | Banditen {pl} :: bandit | bandits
```

The English words, which are on the right side of the two colons, can be ignored, as can the information in braces. Words to the right of the pipe symbol are plural forms and can also be ignored. So what remains is this:

```
Bandit; Räuber
```

These two words are indeed synonyms in German.

Of course not all entries in the dictionary use more than one German word to translate an English word, and even those that do might give misleading results, as in the following example:

```
Bildschirm {m}; überwachen :: monitor
```

Bildschirm and *überwachen* have a different meaning (they even belong to different parts of speech) so they do not belong into the same synonym set.

With this approach, about 28,000 words in about 14,000 synonym sets have been automatically imported into OpenThesaurus.

2.2 Database Structure

The OpenThesaurus website has been implemented using the script language PHP (www.php.net) and MySQL (www.mysql.com), a well-known relational database. Both the script language and the database are available for free. The most essential tables are:

- `words` – lists all words (each word can appear in this table only once)
- `meanings` – lists all meanings
- `word_meanings` – an $n:m$ relation between words and meanings, i.e. a meaning can contain more than one word (which is the common case for a thesaurus) and a word can occur in more than one meaning

2.3 User Interface

Users need to be logged in at the website to get writing access. Users who are not logged in can search the thesaurus data, but they cannot modify anything. When registering, a password is mailed to the user's email address.

The OpenThesaurus homepage enables users to search for words and offers a random search feature. The latter is useful if one wants to improve the thesaurus data but does not know where to start. The database logs how often synonym sets have been displayed. Whenever a new random synonym set is requested by a user, the system selects one that has not been displayed often yet.

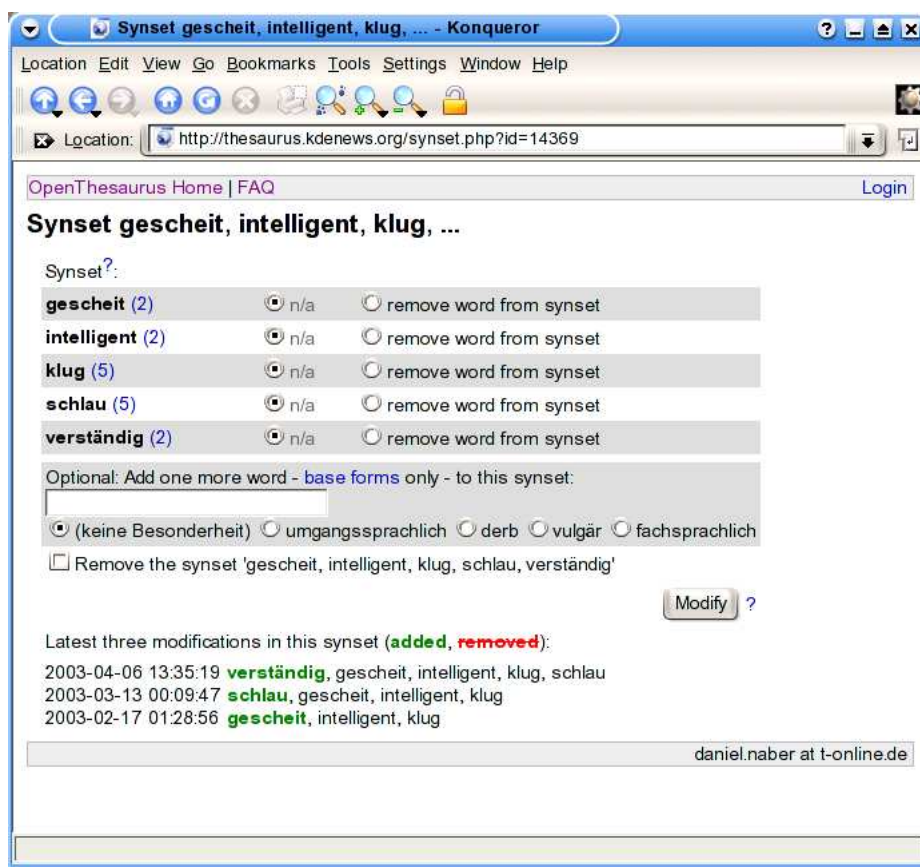


Figure 1: The synonym set of the German word *intelligent*

If a search does not return any matching words, several measures are used to suggest similar words:

- The thesaurus itself only knows base forms, but OpenThesaurus also contains a large database table which maps inflected forms to base forms. For example, a search for *ging* (*walked*) will suggest to search for *gehen* (*walk*).
- All known words are checked for their similarity compared to the input using the Levenshtein algorithm so that, for example, a search for the misspelled word *Quitessenz* will suggest to search for *Quintessenz*.
- The website conducts a substring search. For example, a search for *divers* will suggest *Diversifikation* and *diversifiziert*.

The web page that displays a synonym set offers five main features, as shown in figure 1:

- A new word can be added to the synonym set. Each word can have usage information attached to it (colloquial, vulgar, etc).
- Any word can be removed from the synonym set.

- The complete synonym set can be removed (this will not delete the data but hide it from users).
- The synonym set's superordinate meaning (which is another synonym set) can be defined.
- A list of recent activities by other users for this synset (words added and removed) is displayed.

On an extra page, users can also add new synsets.

2.4 Administration Area

The administrator has access to a special area which contains:

- A list of the users' latest actions, e.g. synonym sets which have been modified, added or removed.
- A list of recently registered users.
- Links to export the thesaurus data as a file for OpenOffice.org, KWord, or a plain text file. Typically the export function will be called automatically once a day.
- A list of recent search queries and their number of matches.
- Several pre-defined database queries, e.g. a search for synonym sets with many/few words, a search for words which appear in more than five synonym sets etc.

In a nutshell, the administration area is useful for quality assurance. Anything users change in the database is logged, so the administrator has full control over all changes.

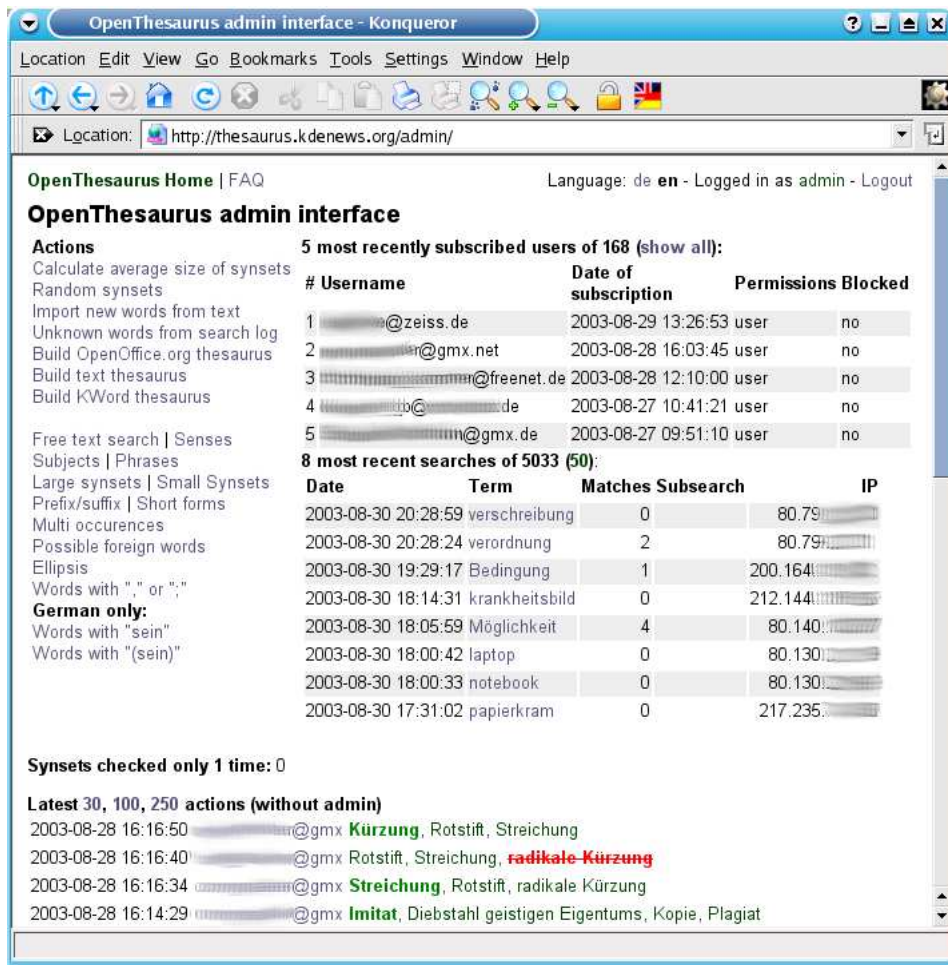


Figure 2: The administrator's interface

2.5 Data Export

The tables which contain the actual thesaurus information – words, meanings, and their relations – are automatically exported using MySQL's `mysqldump` command each night. The file is then compressed and can be downloaded from the website. Users can import it into their local MySQL database to write their own queries that export data. However, modifying the data of a downloaded database dump is not a good idea, as it is difficult to merge it back into the official OpenThesaurus database. Tables that contain detailed information about the user's input, e.g. which user added something into the database at which time, are not exported to protect the users' privacy.

Another nightly export generates thesaurus files which can be used by OpenOffice.org. There is a PHP page in the administration area which builds a list of words and another list which contains the synonyms for each word. Each list is written to a file and then these text files are used as input for Pavel Janík's `Parse_Thes.awk` script which produces the files `th_de_DE.dat` and `th_de_DE.idx`. These can then be used by OpenOffice.org (detailed installation instructions are available in the `README` file in the exported ZIP archive).

2.6 Using OpenThesaurus for Languages other than German

Here is a list of things that need to be done so that another language can use OpenThesaurus to create their own thesaurus:

- The OpenThesaurus PHP files are only available via CVS [OpenThesaurus Project]. They have to be installed in a server's webroot and the database structure needs to be created. The user interface should be translated to the new language using GNU's `i18n` technology (`gettext`). Detailed instructions are available in a separate `README` file. Experience with PHP and MySQL is required prior to setting up such a server.
- The new language project is supposed to find an initial set of data of at least 5,000 words. Starting with less words would be frustrating for users as even many common words will not be included. The initial data needs to be under an Open Source license. This data should preferably be of an OpenThesaurus-like structure, i.e. it should be organized by meanings (*meaning* is another word for *synonym set*). A multilingual dictionary might be a good starting point, as described in 2.1. If nothing else is available, a simple list of words might be used. It is important to invest enough time into finding good and extensive initial data, for example by searching the web for university projects or Open Source projects with similar aims. It does not matter how exactly the initial data is encoded (e.g. as a simple ASCII text file or as a database dump) as long as it can be imported into a relational database. Here is an example that shows initial data which would be useful:
car, auto, automobile, machine
tree, tree diagram
depository financial institution, bank
bank, bank building
- The new language project is supposed to be controlled by an administrator who is a native speaker of that language. His role will be to check the data quality by manually evaluating a random sample of the words in the database. He should also prevent users from adding useless or wrong data into the database by regularly checking all changes made by users.

One of the advantages of using a meaning-based structure is that the synonym sets of two different languages can link to the same meaning, making it possible to automatically build a two language dictionary from the data. For this, the data has to be imported into a joint database and a 1:1 mapping from the meanings of one language to the meanings of the other language needs to be defined.

2.7 Statistics

The OpenThesaurus website was first publically announced on 2003-03-07 on a German OpenOffice.org mailing list. The real-time statistics on the homepage show the progress that has happened since then (fig-

ures as of 2004-06-06):

Number of different words	30,488
Number of synonym sets / meanings	13,039
Added words and new synonym sets:	14,994

The average synonym set currently consists of 2.86 words¹. Each synonym set has been checked at least once to find errors. There are probably some errors left, so the checking will continue. The number of subscribed users is 410. Unfortunately many of these users have subscribed to the site but never actively contribute to it – even though the only reason to subscribe is to be able to contribute, as read-only access is also possible without a subscription.

3 Conclusion

OpenThesaurus has successfully developed the first German Open Source thesaurus. It includes more than 30,000 words and few errors thanks to repeated quality checks.

The rather simple structure can easily be understood by users, whether or not they have a background in linguistics. An estimated 95% of all user contributions are correct and do not need to be corrected or reverted by the administrator. Sadly, only a rather small number of users actively contribute to the project.

The OpenThesaurus website is implemented in PHP, a common programming language and can thus relatively easy be installed on many web servers. The first non-German OpenThesaurus project is the Spanish one, a Polish and a French project is being prepared, others will hopefully follow.

A Document History

2004-06-06	Mention superordinate/subordinate relation. Updating the statistics.
2004-04-04	Fixed some spelling and grammar mistakes
2004-01-10	Removed a misleading example in chapter 2.1
2003-12-04	Some grammar and style fixes, thanks to Christian B. and Peter Reinhart. Added a link to the OpenThesaurus project on SourceForge.
2003-11-17	Fixed some typos, updated figure 1
2003-11-02	Paper published on the OpenThesaurus homepage

References

- [Dict] Frank Richter: *German <-> English Dictionary - TU Chemnitz*,
<http://dict.tu-chemnitz.de/>
- [Duden] *Duden: Die deutsche Rechtschreibung*, Mannheim, 1996
- [GermaNet] <http://www.sfs.uni-tuebingen.de/lsd/>
- [GPL] *GNU General Public License*,
<http://www.gnu.org/copyleft/gpl.html>
- [OpenThesaurus Project] <http://sourceforge.net/projects/openthesaurus/>
- [OpenThesaurus Website] <http://thesaurus.kdenews.org/>
- [WordNet] <http://www.cogsci.princeton.edu/~wn/>
- [Wortschatz Deutsch] <http://wortschatz.uni-leipzig.de/>

¹As a word can appear in more than one synset, this is *not* just the number of different words divided by the number of synsets.