СМЫСЛОВОЕ ВЫРАВНИВАНИЕ, ОСНОВАННОЕ НА ЛИНГВИСТИЧЕСКОЙ МОДЕЛИ, КАК СРЕДСТВО ИНТЕГРАЦИИ НОВОГО ЯЗЫКА В МНОГОЯЗЫЧНУЮ ЛЕКСИКО-СЕМАНТИЧЕСКУЮ БАЗУ ДАННЫХ С ИНТЕРЛИНГВОЙ

Гончарова М. Б. (maria_go@abbyy.com),
Козлова Е. А. (Helen_Koz@abbyy.com),
Пасюков А. В. (Artem_P@abbyy.com),
Гаращук Р. В. (Ruslan_G@abbyy.com),
Селегей В. П. (Vladimir_S@abbyy.com)

ABBYY, Москва, Россия

Ключевые слова: смысловое выравнивание, многоязычные лексико-семантические ресурсы, интеграция новых языков

MODEL-BASED WSA AS MEANS OF NEW LANGUAGE INTEGRATION INTO A MULTILINGUAL LEXICAL-SEMANTIC DATABASE WITH INTERLINGUA

Goncharova M. B. (maria_go@abbyy.com),
Kozlova E. A. (Helen_Koz@abbyy.com),
Pasyukov A. V. (Artem_P@abbyy.com),
Garashchuk R. V. (Ruslan_G@abbyy.com),
Selegey V. P. (Vladimir_S@abbyy.com)

ABBYY, Moscow, Russia

This paper presents a model-based approach to Word Sense Alignment (WSA) applied for new language integration within ABBYY Compreno lexical-semantic database with interlingua. Using the model, i.e. semantic and syntactic compatibility, we perform semantic-syntactic analysis with language-independent structure as a result. With the comprehensive description of core languages at our disposal, we analyze parallel resources, namely, the part of a bilingual dictionary and of a parallel corpus in a source language, and obtain a set of candidate concepts for meanings of a target

language. In this way, we accomplish WSA between the dictionary meanings and the concepts of interlingua. Once the correspondences between the meaning and the concepts of the hierarchy are established, these new meanings can be incorporated into the lexical-semantic database. The integration is fulfilled semi-automatically, i.e. at the final stage the correspondences are to be approved by a linguist; however, the amount of manual work is reduced to minimum.

Key words: word sense alignment, multilingual lexical-semantic resources, new language integration

1. Introduction

In recent years, quick integration of new languages into multilingual lexical-semantic resources (LSR) has been one of the key challenges facing the NLP-community. Despite being time and money consuming venture, the task is nevertheless indispensable for all cross-lingual NLP applications based on semantics. Initially, LSR were mainly expert-built, which required years of manual work. The most well-known and inventory-rich expert-built lexical-semantic database is Princeton WordNet (PWN) and multilingual resources centered around it.

ABBYY Compreno Technology was also created on the basis of a multilingual LSR developed by linguists. The system is centered around interlingua, a hierarchy of language-independent concepts serving as a link between languages and resources, and is based on the model. The term 'model' stands for a full description of semantic and syntactic compatibility of a given meaning [Manicheva et al., 2012]. Therefore, the description is voluminous and requires much effort in terms of manpower and duration. However, the already existing comprehensive description allows to speed up new language integration considerably.

Within the present article, we report on the approach to new language integration hinging on model-based WSA. One of the key implementations of WSA [Matuschek, 2014] is to bring together heterogeneous pieces of information pertaining to a given meaning presented in different LSRs. However, thanks to interlingua and language-independent output of semantic-syntactic analysis, WSA can also be employed for new language integration within interlingua-based systems such as Compreno.

As stated above, we define our approach as model-based. Using the model as a reference point, we perform a semantic-syntactic analysis of a part of the available bilingual resources (bilingual dictionaries and parallel corpora) in a source language that has already been described (in our case, English). Due to universal structure of Compreno LSR, the semantic-syntactic analysis provides a set of candidate language-independent concepts for the meanings of the target language (in our case, German). Once the correspondences between the meaning and the concepts of the hierarchy are established, these new meanings can be incorporated into Compreno lexical-semantic database. The integration is fulfilled semi-automatically, i.e. at the final stage the correspondences are to be approved by a linguist.

The paper is organized as follows. Section 2 presents the existing approaches to WSA and new language integration to LSRs. In Section 3, we concentrate on our background, briefly describing the Compreno language model and how it is used for semantic-syntactic analysis. Section 4 is devoted to the methodology of the present approach. Section 5 introduces the evaluation results. Finally, Section 6 contains our conclusions and illustrates possible further development.

2. Overview

2.1. Approaches to new language integration to LSRs

For the systems based on one core-language, we can distinguish two approaches to integration of new languages [Vossen, 1998]:

- **the merge model** presupposes creating a new hierarchy for the target language with subsequent linking of its nodes with those of the source LSR. This model was mostly used at the early stages of multilingual LSR development [Azarova, 2008; Tufis et al., 2004].
- **the expand model**: The expand model exploits the structure of PWN filling it with the meanings of new languages[Pianta et al., 2002, Robkop et al., 2010; Wang and Bond, 2013]. Being mostly translation-driven, this model relies on various bilingual [Oliver and Climent, 2014; Pradet et al., 2014; Fisher and Sagot, 2008] as well as collaboratively-built (Wikipedia, OmegaWiki, Wiktionary) resources [Pilehvar and Navigli, 2014].

The expand model approach to integration of a new language on the basis of parallel bilingual corpora originates from the presumption that the translations of words in real texts shed light on their semantics [Resnik and Yarowsky, 1997; Mikolov et al., 2013]. There are two strategies for automatic construction of such corpora:

- **by machine translation of sense-tagged corpora** [Oliver and Climent, 2012]
- **by automatic sense tagging of bilingual corpora [**Oliver and Climent, 2014].

The same methods can be applied to interlingua-based LSRs, as our current work demonstrates. As a matter of fact, in our experiments we are using a set of methods associated with the expand model because we process various bilingual resources.

2.2. Approaches to WSA

The primary goal of WSA is to unify the information associated with a given meaning through linking pairs of senses (or, more generally, concepts) from two LSRs, where the members of each pair represent an equivalent meaning [Matuschek, 2014]. There are several approaches commonly used for this task: approaches based on the similarity of textual descriptions of word senses, approaches based on structural properties of LSRs, and a combination of both.

In the framework of **similarity-based approaches,** the meanings are aligned according the similarity glosses, i.e. textual descriptions of word senses. Using this method, Niemann and Gurevych [2011] aligned WordNet to Wikipedia, while Meyer and Gurevych [2011] aligned WordNet to Wiktionary, calculating cosine or personalized page rank (PPR) similarity [Agirre and Soroa, 2009] and using simple machine learning techniques for sense classification. Later on, the same approach was chosen for cross-lingual alignment between WordNet and the German part of OmegaWiki [Gurevych et al., 2012], with machine translation as an intermediate component.

Within **graph-based approaches,** structural properties of LSRs are the main criteria for linking senses. Thus, Ponzetto and Navigli [2009] built subgraphs of WordNet for each Wikipedia category to align WordNet synsets and Wikipedia categories. Alternatively, Matuschek and Gurevych [2013] apply a kind of graph-based approach, Dijkstra-WSA, to align different resources (WordNet-OmegaWiki, WordNet-Wiktionary, GermaNet-Wiktionary and WordNet-Wikipedia) using the shortest path lengths.

Currently, a **hybrid approach** is also in use, where distances between senses in the graph representations of LSRs are taken into account along with gloss similarities [Matuschek and Gurevych, 2014].

As we have already pointed out, in this paper we present a model-based approach to WSA. We align the German meanings in a bilingual dictionary with the concepts of the SH through parsing of a bilingual German-English dictionary and a parallel German-English corpus. A more detailed description of Compreno semantic model and the process of semantic-syntactic analysis will help to understand how this approach was developed.

3. Background

3.1. Compreno Description

The core of Compreno linguistic model is a universal **Semantic Hierarchy** (SH) based on interlingua. **Interlingua**, a language-independent level of concepts, serves as a link between different languages (Fig.1). At present, the SH contains 141,342 concepts. The description of Russian and English are almost complete; the integration of German is well underway; French, Chinese and Spanish are at the initial stage of description.

The SH is a hierarchical tree organized according to hyper-hyponymy relations. For each node only one direct ascendant is possible. The nodes of this tree are called **Semantic Classes** (SC) and represent language-independent "meanings". An SC contains **Lexical Classes** (LC) that represent language-specific meanings. In their turn, the LCs contain words, i.e. language-specific lexemes. It is worth mentioning that the structure of Compreno SH is POS-independent; consequently, lexemes belonging to different parts of speech can be comprised within a single meaning, depending on the model of the branch. The meanings within SC are either synonyms, or antonyms, and differ by a set of **semantemes**, units of universal semantic information, e.g. <<PolarityPlus>>, <<PolarityMinus>>, <<Bookish>>, <<Special>>,

<<Elevated>>, etc. Semantemes also encode more specific semantic relations that are not explicitly reflected in the structure of the SH, for instance, <<Part>>, <<Whole>>, or <<SingulativePortion>>.

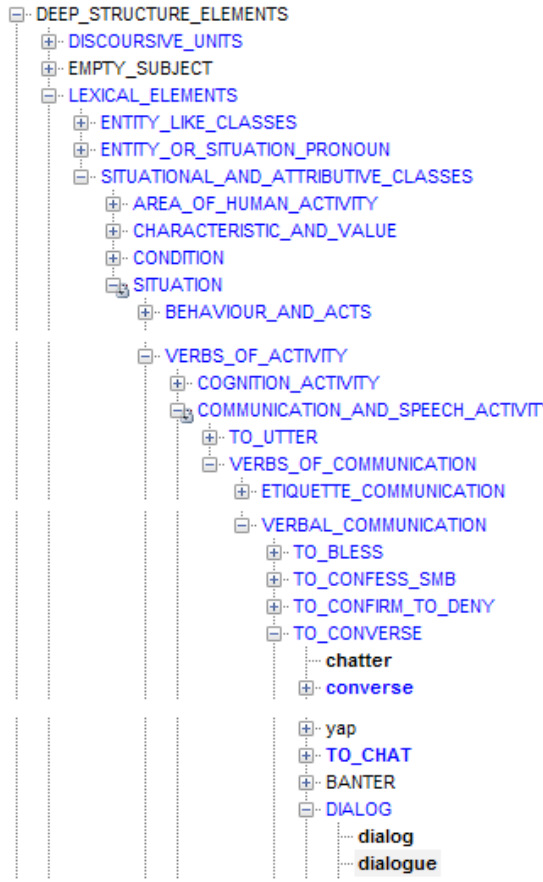

Figure 1. A Fragment of the Semantic Hierarchy

Since from the very beginning the system has been conceived as a multilingual database aimed at machine translation, each meaning is provided with a morphological, lexical semantic, and syntactic description.

The key feature of Compreno technology is that each concept and each meaning in the SH has **a semantic and syntactic model**, i.e. semantic and syntactic compatibility, which is inherited from the higher levels of the SH. Semantic compatibility is described by means of language-independent **semantic slots** (more than 300), which, to some extent, correlate with semantic valencies in L. Tesnière's dependency grammar theory [Tesnière, 1959], with deep cases in Ch. Fillmore's case grammar theory [Fillmore, 1968]. Syntactic compatibility, on the other hand, is described with the help

of **syntactic slots** that represent language-specific realizations of semantic slots. Syntactic characteristics of a meaning are unified within a **syntactic paradigm,** which includes a **universal syntactic paradigm** (syntactic characteristics of different POS) and a **lexical syntactic paradigm** (syntactic properties of a given meaning). The description also comprises non-tree syntax (regulates conjunction links, structural control, pronoun resolution, etc.), and analysis rules (preserve/extract universally-relevant bits of grammatical meaning, such as Tense and Modality of verbs, or the Number of substantives).

The set of semantic and syntactic properties, coupled with unsupervised machine learning through the use of an automatically labeled corpus, allows to deal with Word Sense Disambiguation (WSD) for a concrete language. Since Compreno has been conceived as a multilingual model, it also provides features for **treatment of cross-language asymmetry**. Cross-language hyperonym-hyponym asymmetry is neutralized by the ability to choose translation equivalents from both parent and child SCs [Manicheva, 2012]. Lexical gaps are filled with multiword expressions (terms and idioms). Within our system, **terms** are not just concepts relating to a certain domain. They are always multiword and are situated right under the SC of their root nodes, inheriting all their properties. In linguistics, **idioms** are usually presumed to be figures of speech that contradict the principle of compositionality. This principle states that the meaning of a whole should be constructed from the meanings of the parts that make up the whole). In the framework of our system, idioms are positioned according to the meaning of the whole expression.

3.2. Stages of text analysis

The distinctive feature of the approach presented in the article is full semantic-syntactic analysis of bilingual resources. We perform automatic sense-tagging of the English part of the parallel corpus by means of ABBYY Compreno parser. An important aspect of Compreno parsing technology is that syntactic and semantic disambiguation are processed in parallel from the very beginning (in contrast to the architecture more usual for the NLP systems where the semantic analysis follows the syntactic one [Anisimovich et al., 2012]).

The analysis is performed in several stages (Fig. 2). Semantic ambiguity remains unresolved as long as possible. The first stage is **lexical-morphologic**, where we consider all possible LCs for a given lexeme with all possible morphological meanings. At the stage of **syntactic analysis,** we build a syntactic graph. Initially, the edges of the graph are labeled with all possible syntactic and semantic relations, as well as grammatical properties. Gradually, incompatible meanings are eliminated. At the same time, the system checks for non-tree relations, if any. As a result of the filtration of incompatible meanings, we obtain one or several semantic-syntactic structures, each of which has the right to exist due to semantic-syntactic homonymy. **The final semantic-syntactic structure** is chosen according to statistical evaluation [Zuev, 2013]. Finally, **the universal semantic structure** with semantic relations and meanings is built through removal of all the language-specific information (surface slots, LCs and grammatical meanings).

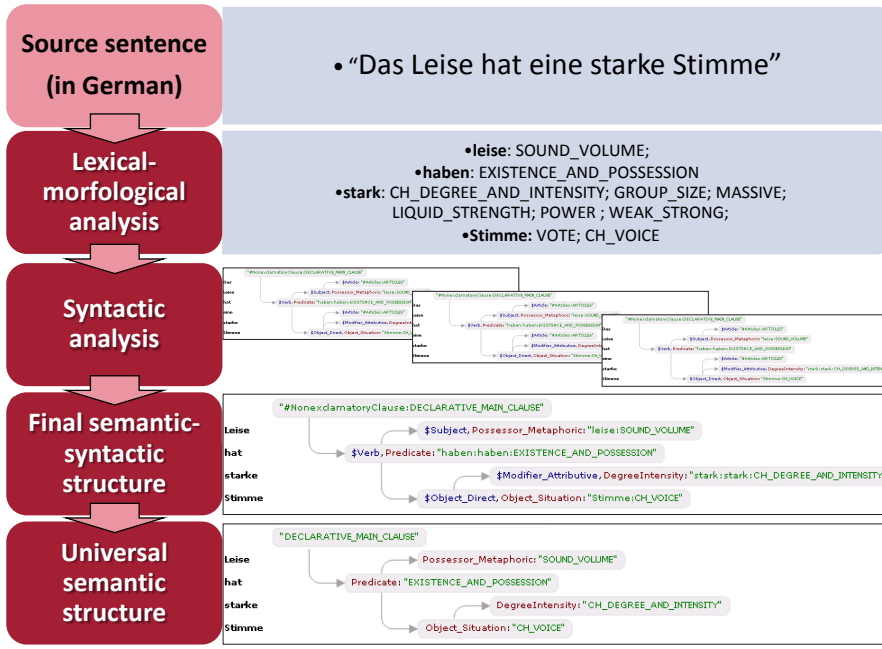

Figure 2. Stages of the Semantic-Syntactic Analysis

4. Methodology

4.1. Parallel corpus processing and statistical data retrieval

We carry out word alignment of a large parallel English-German corpus (10,250,572 sentences). Matching is accomplished using the Hungarian method for constructing a maximum weighted bipartite graph matching [Kuhn, 2010].

4.2. Statistical data filtration

As a result of word alignment, we obtain a list of pairs 'German lexeme—English translation variant'. Then, the English part of the corpus is parsed, and we obtain a list of pairs 'German lexeme—SC', with a frequency score for each correspondence. This list is called henceforth statistical data, or statistics. We filter out low frequency results (1/10000 of the maximum value) for each lexeme. At the next stage, it is important to distribute the resulting pairs across the meanings of the dictionary[1] entry.

[1] PONS Wörterbuch Englisch Premium.
Number of headwords: 98093. Number of entries: 97946. Version: 1.0 (01.11.2011)
Source: PONS Wörterbuch Englisch Premium. Based on PONS dictionary contents www.pons.de
© PONS GmbH, Stuttgart 2011

For these purposes, we perform semantic-syntactic analysis of the entry. In fact, the main principle that underlies our approach is quite simple: we obtain parallel corpus data and data from the dictionary entry and intersect the two sets (see Fig. 3).

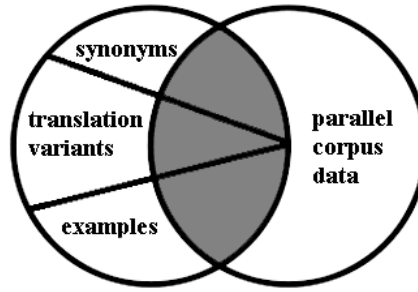

Figure 3. Venn diagram: Intersection of Data Received from the Parallel Corpus and the Dictionary

The principle described above is realized by means of a heuristically based algorithm (Fig. 4). In order to assign a given SC to the meaning of the entry, the program takes a pair 'German lexeme-SC' and decides whether it can be added through semantic analysis of the entry. We have a special environment with an integrated dictionary, where candidate SCs can be added as links to the SH. Once the link is validated by synonym, translation, or example, it is marked as a good link for the given meaning. At every stage of the analysis, the POS of the German lexeme is compared to that of the lexeme in the candidate SC. If the SC contains a lexeme belonging to the same POS, the link is marked as good. If not, the link is retained only if there are no other results.

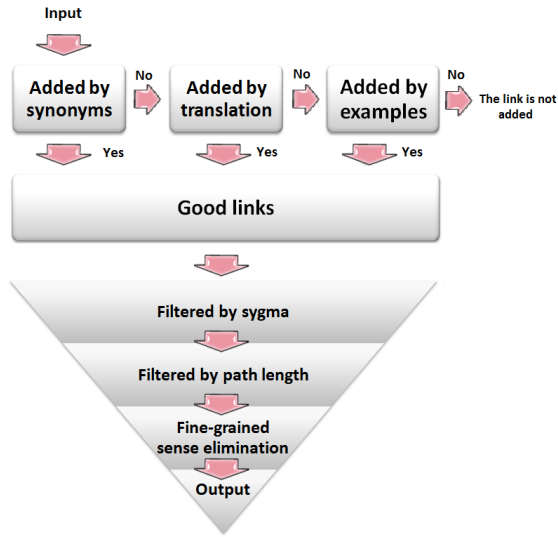

Fig. 4. The Algorithm of Adding a Candidate SC

German synonyms. As the language integration technology is semi-automatic, the German lexical semantic database is filled gradually. Consequently, we can use the analysis of those German words, which are still being added into the SH.

It often occurs that a hyperonym is indicated in brackets instead of a synonym, so we take into account parent-child relations. For example, the second meaning of *'zünden'* (Table 1) is explained through a hyperonym *'wirken'* (SC *'CH_POWER_AND_EFFECT'*). In this case, we retain SC *'TO_BLIGHT_AS_TO_AFFECT'* as a possible candidate because the class is a descendent of the SC *'CH_POWER_AND_EFFECT'*.

Table 1. Parallel Corpus, Dictionary Entry Data, and the Output for *'zünden'*

Parallel corpus data		Dictionary data	CS-candidates	Added by
{ARDENT}	114	*vt* 1) *TECH* ▪ etw zünden to fire sth *spec*	TO_BURN TO_ACTIVATE	**translation** *fire*
{DETONATION}	36			
{EMOTIONAL_STATE}	15			
{FIRE_AS_EMERGENCY}	117			
{FIRE_SHOOTING}	326			
{FIRE}	138	2) *(wirken)* to kindle enthusiasm	TO_BLIGHT_AS_TO_AFFECT	**synonym** *wirken*
{IGNITION}	21			
{INSIGHT_INTO}	30			
{TO_ACTIVATE}	225			
{TO_BLIGHT_AS_TO_AFFECT}	38	3) example: ▶ hat es bei dir endlich gezündet? — have you cottoned on? *fam*, *BRIT a.* has the penny dropped? *fam*	TO_UNDERSTAND: INSIGHT_INTO	**example** *have you cottoned on?*
{TO_BURN}	202			
{TO_COIN}	28			
{TO_CROSS_OUT}	39			
{TO_EVOKE}	21			
{TO_RETIRE}	182			
{TO_SET_THE_HOOK}	14			
{TO_TREAT_WITH_FIRE}	90			

Translation variants. When we perform semantic-syntactic analysis of the translation variants, we take into consideration all possible semantic-syntactic structures of the target language. E.g. the English translaton*'to fire smth'* for the first meaning (Table 1) gives us the SCs from statistical data*'TO_BURN'* and *'TO_ACTIVATE'* as candidate SCs.

Examples. We add the SCs from the statistics that coincide with the SCs derived from the semantic-syntactic analysis of the example, with only the best structure chosen. Thus, the SC *'INSIGHT_INTO'*, appearing in statistics, is confirmed by the example (the principle of parent-child relations is relevant here as well):

(1) #[[**have**] [**you** "#pronoun_personal:#pronoun_personal:PRONOUN_BEING"] **cottoned** "cotton_on:TO_UNDERSTAND" [**on**]]?

In this way, for each meaning we obtain a set of "good links", which undergoes a number of filtrations afterwards.

Filter 1: We calculate the standard deviation where the maximum value is used instead of the mean value (Fig. 5). Thus we determine the threshold value of frequency for every meaning. All the links that lie below the threshold are filtered out.

Filter 2. In order to reduce the number of irrelevant links, we introduce additional scores which reflect the degree of affinity with the units added by other elements of the dictionary entry for each link (Table 2). The score is calculated as a ratio between the sum of maximum coinciding path lengths and the length of a given link. To obtain the maximum coinciding path length we compare the path of a link (for instance, *'PERMISSION'*) added by one element of the entry (*'permit'*) with the links added by other elements (*'Ausweis', 'pass'*). In the example, both *'Ausweis'* and *'permit'* have the maximum coinciding path length of 6 (taken from the root of the SH). All candidate SCs with a score below 0,75 of the maximum score for a given meaning are filtered out.

Table 2. The Path Length Filter

	Ausweis	permit	pass
Legitimation *:f (geh)* **2)** *(Ausweis)* permit, pass	DEEP_STRUCTURE_ ELEMENTS : LEXICAL_ELEMENTS : ENTITY_LIKE_CLASSES : ENTITY : INFORMATION_AND_ SOCIAL_OBJECTS : CREATIVE_WORK : MATERIAL_CREATIVE_ WORK : TEXT_OBJECTS_AND_ DOCUMENTS : DOCUMENT : CERTIFICATE : Ausweis	DEEP_STRUCTURE_ELEMENTS : LEXICAL_ELEMENTS : ENTITY_LIKE_CLASSES : ENTITY : INFORMATION_AND_SOCIAL_ OBJECTS : CREATIVE_WORK : RESULTS_OF_SPEECH_MENTAL _ACTIVITY : RESULTS_OF_GIVING_INFORMA TION_AND_SPEECH_ACTIVITY : PERMISSION_PROHIBITION : PERMISSION : permit	DEEP_STRUCTURE_ELEMENTS : LEXICAL_ELEMENTS : ENTITY_LIKE_CLASSES : ENTITY : INFORMATION_AND_SOCIAL_ OBJECTS : CREATIVE_WORK : MATERIAL_CREATIVE_WORK : TEXT_OBJECTS_AND_ DOCUMENTS : DOCUMENT : WRITTEN_PERMISSION_AS_ LEGAL_DOCUMENT : PASS : pass
			DEEP_STRUCTURE_ELEMENTS : LEXICAL_ELEMENTS : SITUATIONAL_AND_ATTRIBUTIVE_ CLASSES : SITUATION : EXISTENCE_AND_POSSESSION : GIVE_GET_TAKE_AWAY : TO_GIVE : TO_GIVE_TO : pass : pass
			DEEP_STRUCTURE_ELEMENTS : LEXICAL_ELEMENTS : SITUATIONAL_AND_ATTRIBUTIVE_ CLASSES : SITUATION : POSITION_AND_MOTION : MOTION : TO_GO_AND_TRANSFER : pass : pass

Filter 3: Fine-grained sense elimination is applied when we obtain both parent and direct child SCs as candidate SCs for a given meaning. As we have already pointed out, Compreno technology has its own mechanisms for treating cross-lingual asymmetry in hyponym-hyperonym relations with certain classes marked as "transparent". Consequently, the most general concept is retained at this stage.

4.3. Additional Semantic-Syntactic Analysis of a Dictionary Entry

Additional semantic syntactic analysis of a dictionary entry is applied to multiword expressions and to dictionary meanings without candidate SCs. We extract German equivalents of English multiword expressions, irrespectively of whether their German equivalents are multiword or not (see Table 3). For definitions of terms and idioms within our system see Section 3.

Table 3. Treatment of Multiword Expressions

	English	Semantic Class	German
Terms	naval officer	NAVAL_OFFICER	Marineoffizer
	fashion designer	FASHION_DESIGNER	Modemacher
Idioms	getting rid of	GET_RID_OF	Abwicklung
	polar circle	POLAR_CIRCLE	Polarkreis

Additional semantic analysis of the entry allows us to assign links even when parallel corpus data and dictionary data do not coincide.

5. Evaluation and discussion

As a result of the semi-automatic German language integration, 121,852 meanings of 92,985 entries from PONS dictionary were assigned candidate links to SCs (Table 4).

Table 4. Number of Entries and Meanings in the Dictionary

	Meanings	Entries
Nominal	82,854	7,1808
Verbal	19,241	9,173
Adjectival	13,605	10,427
Adverbial	4,124	3,066

To evaluate the effectiveness of the method described above we have taken a random sample of 400 German lexemes from the dictionary. We established a benchmark by manually assigning correct SCs to each of these lexemes. Subsequently, we took the results of our integration method for the sample and compared the two sets. The following measures were computed:

- **precision,** that is, the percentage of relevant SCs retrieved with respect to the number of retrieved SC-candidates;
- **recall,** that is, the percentage of relevant SCs retrieved with respect to the total number of relevant manually assigned SCs.
- **F-score,** calculated according to the formula:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

The results are presented in Table 5.

Table 5. Evaluation Results

	Overall	Monosemous words	Polysemous words
Precision	0.60	0.63	0.52
Recall	0.80	0.82	0.76
F-score	0.69	0.71	0.61

As can be seen from Table 5, we have achieved good results in terms of recall, both for polysemous and monosemous words. Since our SC-candidates are supposed to be later approved by a linguist, precision was not our primary goal. However, precision results can be still improved, which we are planning to do in the framework of our future work. In order to reduce the number of irrelevant SC-candidates, we intend to expand our use of multilingual resources.

It is common knowledge that both dictionary content and dictionary word sense distinction have a rather arbitrary and subjective nature, so it is risky to use one dictionary as a reference point; however, in our case it is checked and enriched through parallel corpus data. For sense distinction, we rely mostly on the structure of our SH, as the meanings are determined by the model and are verified by machine translation of real texts. As a result we get coarse-grained sense distinction based on empirical data.

6. Conclusion

In this article we presented a model-based approach to WSA which we use to integrate a new language (German) into Compreno lexical-semantic database with interlingua. The approach involves semantic-syntactic analysis of the English part of a parallel corpus and a bilingual dictionary. The resulting language-independent structure enables us to deal effectively with cross-language WSD and to carry out cross-language WSA of German meanings with the concepts of the hierarchy.

This approach has the following advantages. Comprehensive description of English within the system and a large-scale parallel corpus enables us to obtain a set of candidate semantic classes for practically every meaning in the German-English dictionary. There is no discrepancy between the results obtained for monosemous and for polysemous words. As the Compreno Semantic Hierarchy does not segregate words by parts of speech, we are able to process all POS in one iteration.

The interlingua level of the hierarchy can also be used for a variety of purposes, besides integration of new languages. For example, it can be applied as an intermediate component for alignment of other resources. Specifically, we are planning integration of different multilingual resources to improve our precision results. Complete replicability of the present experiment is possible within Compreno framework; replicability of the model-based approach is possible within any system with deep semantic analysis.

References

1. *Agirre E., Soroa A.* (2009), Personalizing PageRank for Word Sense Disambiguation, Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-2009), Athens, pp. 33–41.
2. *Anisimovich K. V., Druzhkin K. Ju., Minlos F. R., Petrova M. A., Selegey V. P., Zuev K. A.* (2012), Syntactic and semantic parser based on ABBYY Compreno linguistic technologies, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"], Bekasovo, pp. 90–103.
3. *Azarova I.* (2008), RussNet as a computer lexicon for Russian, Proceedings of the 16th International Conference Intelligent Information Systems, Zakopane, pp. 341–350.
4. *Fillmore Ch.* (1968),The case for case, in E. Bach, R. Harms (eds.), Universals in linguistic theory, New York, Holt, Rinehart and Winstonm, pp. 1–90.
5. *Gurevych I., Eckle-Kohler J., Hartmann S., Matuschek M., Meyer Ch. M., Wirth Ch.* (2012), UBY—A Large-Scale Unified Lexical-Semantic Resource Based on LMF, Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL'12), Avignon, pp. 580–590.
6. *Kuhn H. W.* (2010), The Hungarian method for the assignment problem. In 50 Years of Integer Programming 1958–2008, pp. 29–47.
7. *Manicheva E. S., Petrova M. A., Kozlova E. A., Popova T. V.* (2012), Compreno Semantic Model as Integral Framework for Multilingual Lexical Database, Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon (CogALex-III), COLING 2012, Mumbai, 215–229.
8. *Meyer Ch. M., Gurevych I.* (2011), What Psycholinguists Know About Chemistry: Aligning Wiktionary and WordNet for Increased Domain Coverage, Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP), Chiang Mai, pp. 883–892.
9. *Matuschek M., Gurevych I.* (2013), Dijkstra-wsa: A graph-based approach to word sense alignment. Transactions of the Association for Computational Linguistics (TACL), pp. 151–164.
10. *Matuschek M., Gurevych I.* (2014), High Performance Word Sense Alignment by Joint Modeling of Sense Distance and Gloss Similarity. Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics, Dublin, pp. 245–256.
11. *Matuschek M.* (2014), Word Sense Alignment of Lexical Resources, Ph. D. thesis, Technische Universität Darmstadt.
12. *Mikolov T., Le Q. V., Sutskever I.* (2013), Exploiting similarities among languages for machine translation, Computation and Language Archive, abs/1309.4168.
13. *Niemann, E., Gurevych I.* (2011), The People's Web meets Linguistic Knowledge: Automatic Sense Alignment of Wikipedia and WordNet, Proceedings of the 9th International Conference on Computational Semantics, Oxford, pp. 205–214.
14. *Oliver A.* (2014), WN-Toolkit: Automatic generation of WordNets following the expand model, Proceedings of the 7th Global WordNet Conference, Tartu, pp. 7–15.

15. *Oliver A., Climent S.* (2014), Automatic creation of wordnets from parallel corpora, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, pp. 1112–1116.
16. *Oliver A., Climent S.* (2012), Building wordnets by machine translation of sense tagged corpora, Proceedings of the Global WordNet Conference, Matsue, pp. 232–240.
17. *Pianta E., Bentivogli L., Girardi Ch.* (2002), Multiwordnet: developing an aligned multilingual database, Proceedings of the First International Conference on Global WordNet, Mysore, pp. 293–302.
18. *Pilehvar M. T., Navigli R.* (2014), A Robust Approach to Aligning Heterogeneous Lexical Resources. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014), Baltimore, pp. 468–478.
19. *Ponzetto S. P., Navigli R.* (2009), Large-Scale Taxonomy Mapping for Restructuring and Integrating Wikipedia, Proc. of the 21st International Joint Conference on Artificial Intelligence (IJCAI 2009), Pasadena, pp. 2083–2088.
20. *Pradet Q., Chalendar G. de, Desormeaux J. B.* (2014), Wonef, an improved, expanded and evaluated automatic french translation of wordnet. Proceedings of the 7th Global WordNetConference, Tartu, pp. 32–40.
21. *Resnik Ph., Yarowsky D.* (1997), A perspective on word sense disambiguation methods and their evaluation. Proceedings of the ACL-SIGLEX Workshop «Tagging Text with Lexical Semantics: Why, What, and How?», Washington, DC, pp. 79–86.
22. *Robkop K., Thoongsup S., Charoenpron Th., Sornlertlamvanich V., Isahara H.* (2010), WNMS: Connecting Distributed Wordnet in the Case of Asian WordNet, Proceedings of the 5th International Conference of the Global WordNet Association (GWC 2010), Mumbai.
23. *Sagot B., Fišer D.* (2008), Building a free French wordnet from multilingual resources, Proceedings of the Ontolex 2008, Marrakech, pp. 14–19.
24. *Tiedemann J.* (2011), Bitext Alignment, Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, United States.
25. *Tufis D., Ion R., Barbu E., Barbu V.* (2004), Cross-Lingual Validation of Multilingual Wordnets. Proceedings of the Second Global WordNet Conference, Brno, pp. 332–340.
26. *Vossen P.* (1998), EuroWordNet: a multilingual database with lexical semantic networks for EuropeanLanguages. Kluwer, Dordrecht.
27. *Wang Sh., Bond F.* (2013), Building the chinese open wordnet (cow): Starting from core synsets, Proceedings of the 11th Workshop on Asian Language Resources, a Workshop at IJCNLP-2013, Nagoya, pp. 10–18.
28. *Zuev K. A., Indenbom M. E., Judina M. V.* (2013), Statistical machine translation with linguistic language model, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"], Bekasovo, vol. 2, pp. 164–172.