Developing an interlingual translation lexicon using WordNets and Grammatical Framework

Shafqat Mumtaz Virk
University of Gothenburg,
University of Eng. & Tech. Lahore
virk.shafqat@gmail.com

K.V.S. Prasad
Chalmers University of Technology
prasad@chalmers.se

Aarne Ranta
University of Gothenburg
aarne@chalmers.se

Krasimir Angelov
University of Gothenburg
krasimir@chalmers.se

Abstract

The Grammatical Framework (GF) offers perfect translation between controlled subsets of natural languages. E.g., an abstract syntax for a set of sentences in school mathematics is the interlingua between the corresponding sentences in English and Hindi, say. GF "resource grammars" specify how to say something in English or Hindi; these are re-used with "application grammars" that specify what can be said (mathematics, tourist phrases, etc.). More recent robust parsing and parse-tree disambiguation allow GF to parse arbitrary English text. We report here an experiment to linearise the resulting tree directly to other languages (e.g. Hindi, German, etc.), i.e., we use a language-independent resource grammar as the interlingua. We focus particularly on the last part of the translation system, the interlingual lexicon and word sense disambiguation (WSD). We improved the quality of the wide coverage interlingual translation lexicon by using the Princeton and Universal WordNet data. We then integrated an existing WSD tool and replaced the usual GF style lexicons, which give one target word per source word, by the WordNet based lexicons. These new lexicons and WSD improve the quality of translation in most cases, as we show by examples. Both WordNets and WSD in general are well known, but this is the first use of these tools with GF.

1 Introduction

1.1 Translation via an interlingua

Interlingual translation scales easily up to a large number of languages. Google translate, for example, deals with all pairs of 60 languages mostly by using English as a pivot language. In this way, it can do with just 2 * 59 = 118 sets of bilingual training data, instead of 60 * 59 = 3540 sets. It would be hard to collect and maintain so many pairs, and in many cases, there is very little data to be found.

The roots of an inter-lingua are perhaps in the medieval idea of a universal grammar (Lyons, 1968), in which a universal representation of meaning can be expressed. Translating via this interlingua then also means that meaning is conserved in going from the source to the target language. In recent decades, this idea appears in (Curry, 1961) where the interlingua is called tectogrammar, in the Rosetta project (Rosetta, 1994), building on the semantic models of (Montague, 1974), and in the UNL (Universal Networking Language) project.

Incidentally, interlingua is also the heart of modern compiler technology. For instance, the GNU Compiler Collection (Stallman, 2001) uses a shared tree representation to factor out the majority of compilation phases between a large number of source and target languages. Compiler writers save work, and semantics is preserved by design. A compiler, then, is built as a pipeline with **parsing** from a source language to an **abstract syntax tree**, which is analyzed and optimized in the language-independent phases, and finally **linearized** to a target language.

This work is licenced under a Creative Commons Attribution 4.0 International License. Page numbers and proceedings footer are added by the organizers. License details: http://creativecommons.org/licenses/by/4.0/

It is easy to see an analogy between this pipeline and the way a human language translator could work. But how to make it real? How to scale up to the full size of natural languages?

1.2 WordNets

In current machine translation research, interlingual methods are marginal, despite the wide use of pivot languages in systems like Google translate. Closest to the mainstream perhaps is the development of linked WordNets. The original Princeton Wordnet for English (Miller, 1995) defines a set of word senses, which many other wordnets map to other languages. Implementations of this idea are Finnish (Lindén and Carlson., 2010) and Hindi (Hindi-WordNet, 2012).

In the linked Wordnet approach, the Princeton WordNet senses work as an interlingua, albeit only on the level of the lexicon. (Lindén and Carlson., 2010) give strong arguments why in fact this is a good way to go, despite the often emphasized fact that different languages divide the world in different ways, so that the senses of their word don't map one to one. The evidence from the English-Finnish case shows that 80% of the mappings are one-to-one and un-problematic. As this part of the lexicon can be easily reused, linguists and system builders can concentrate their effort on the remaining 20%.

The Universal WordNet (de Melo and Weikum, 2009) works on the same lines. Building on the Princeton WordNet, it populates the mappings to over 200 different languages by collecting data from different sources (such as the Wikipedia) and using supervised machine learning techniques to propagate the knowledge and infer more of it. What makes it a particularly interesting resource is that it is freely available under the most liberal licenses, as is the original Princeton WordNet,

1.3 GF

Grammatical Framework (GF)(Ranta, 2004) is a grammar formalism tool based on Martin Löf's type theory (Martin-Löf, 1982). It can be seen as a tool to build interlingua based translation systems. GF works like a compiler: the source language is parsed to an abstract syntax tree, which is then linearized to the target language. The parsing and linearization component are defined by using Parallel Multiple Context-Free Grammars (PMCFG, (Seki et al., 1991), (Ljunglöf, 2004)), which give GF an expressive power between mildly and fully context-sensitive grammars. Thus GF can easily handle with language-specific variations in morphology, word order, and discontinuous constituents, while maintaining a shared abstract syntax.

Historically, the main use of GF has been in controlled language implementations, e.g., (Ranta and Angelov, 2010; Angelov and Enache, 2010; Ranta et al., 2012) and natural language generation, e.g., (Dymetman et al., 2000), both applied in multilingual settings with up to 15 parallel languages. In recent years, the coverage of GF grammars and the processing performance has enabled open-domain tasks such as treebank parsing (Angelov, 2011) and hybrid translation of patents (Enache et al., 2012). The general purpose Resource Grammar Library (RGL)(Ranta, 2011) has grown to 30 languages. It includes the major European languages, South Asian languages like Hindi/Urdu (Prasad and Shafqat, 2012), Nepali and Punjabi (Shafqat et al., 2011), the Southeast Asian language Thai, and Japanese and Chinese.

However, GF has yet not been exploited for arbitrary text parsing and translation. To do this, we have to meet several challenges: robust parsing, parse-tree disambiguation, word sense disambiguation, and development of a wide-coverage interlingual translation lexicon. This paper focuses on the latter two. We report first a method of using the WordNets (Princeton and Universal) to build an interlingual full-form, multiple sense translation lexicon. Then, we show how these lexicons together with a word sense disambiguation tool can be plugged in a translation pipeline. Finally, we describe an experimental setup and give many examples to highlight the effects of this work.

1.4 South Asian languages

While the work described here can apply to any language, it is particularly interesting for South Asian languages. In these languages, statistical tools do not have much bilingual training data to work on, so Google translate and similar tools are not as useful as they are with better resourced languages. At the same time, there is an urgent and widely recognised need for translations from English to the various languages of South Asia. Fortunately, word nets are being built for many of them, so that the techniques described here can be applied.

2 From Universal WordNet to a GF Lexicon

The original Princeton WordNet (Miller, 1995) defines a set of word senses, and the Universal WordNet (de Melo and Weikum, 2009) maps them to different languages. In this multilingual scenario, the Princeton WordNet senses can be seen as an abstract representation, while the Universal WordNet mappings can be seen as concrete representation of those senses in different languages. GF grammars use very much the same technique of one common abstract and multiple parallel concrete representations to achieve multilingualism. Due to this compatibility, it is easy to build a multilingual GF lexicon using data from those two resources (i.e. Princeton and Universal WordNets). This section briefly describes the experiment we did to build one abstract and multiple concrete GF lexicons for a number of languages including German, French, Finnish, Swedish, Hindi, and Bulgarian. The method is very general, so can be used to build a similar lexicon for any other language for which data is available in the Universal WordNet.

2.1 GF Abstract Lexicon

The Princeton WordNet data is distributed in the form of different database files. For each of the four lexical categories (i.e. noun, verb, adjective, and adverb), two files named 'index.pos' and 'data.pos' are provided, where 'pos' is noun, verb, adj and adv. Each of the 'index.pos' files contains all words, including synonyms of the words, found in the corresponding part of speech category. Each of the 'data.pos' files contains information about unique senses belonging to the corresponding part of speech category. For our purposes, there were two possible choices to build an abstract representation of the lexicon:

1. To include all words of the four lexical categories, and also their synonyms (i.e. to build the lexicon from 'index.pos' files)
2. To include only unique senses of the four categories with one word per sense, but not the synonyms (i.e. to build the lexicon from the data.pos' files)

To better understand this difference, consider the words 'brother' and 'buddy'. The word 'brother' has five senses with sense offsets '08111676', '08112052', '08112961', '08112265' and '08111905' in the Princeton WordNet 1.7.1[1], while the word 'buddy' has only one sense with the sense offset '08112961'. Choosing option (1) means that we have to include the following entries in our abstract lexicon.

```
brother_08111676_N
brother_08112052_N
brother_08112961_N
brother_08112265_N
brother_08111905_N
buddy_08112961_N
```

We can see that the sense with the offset '08112961' is duplicated in the lexicon: once with the lemma 'brother' and then with the lemma 'buddy'. However, if we choose option (2), we end up with the following entries.

[1] We choose WordNet 1.7.1, because the word sense disambiguator that we are using in our translation pipeline is based on WordNet 1.7.1

```
brother_08111676_N
brother_08112052_N
brother_08112265_N
brother_08111905_N
buddy_08112961_N
```

Since the file 'data.noun' lists the unique senses rather than the words, their will be no duplication of the senses. However, the choice has an obvious effect on the lexicon coverage, and depending on whether we want to use it as a parsing or as a linearization lexicon, the choice becomes critical. Currently, we choose option (2) for the following two reasons:

1. The Universal WordNet provides mappings for synsets (i.e. unique senses) but not for the individual synonyms of the synsets. If we choose option (1), as mentioned previously, we have to list all synonyms in our abstract representation. But, as translations are available only for synsets, we have to put the same translation against each of the synonyms of the synset in our concrete representations. This will not gain us anything (as long as we use these lexicon as linearization lexicons), but will increase the size of the lexicon and hence may have reduce the processing speed of the translation system.
2. At the current stage of our experiments we are using these lexicons as linearization lexicons, so one translation of each unique sense is enough.

Our abstract GF lexicon covers 91516 synsets out of around 111,273 synsets in the WordNet 1.7.1. We exclude some of the synsets with multi-word lemmas. We consider them more of a syntactic category rather than a lexical category, and hence deal with them at the syntax level. Here, we give a small segment of our abstract GF lexicon.

```
abstract LinkedDictAbs = Cat ** {
  fun consecutive_01624944_A  : A ;
  fun consequently_00061939_Adv : Adv ;
  fun conservation_06171333_N : N ;
  fun conspire_00562077_V : V ;
  fun sing_01362553_V2  : V2 ;
  ........
 }
```

The first line in the above given code states that the module 'LinkedDictAbs' is an abstract representation (note the keyword 'abstract'). This module extends (achieved by '**' operator) another module labeled 'Cat[2]' which, in this case, has definitions for the morphological categories 'A', 'Adv', 'N' and 'V'. These categories correspond to the 'adjective', 'adverb', 'noun', and 'verb' categories in the WordNet respectively. However, note that in GF resource grammars we have a fine-grained morphological division for verbs. We sub-categorize them according to their valencies i.e 'V' is for intransitive, and 'V2' for transitive verbs. We refer to (Bringert et al., 2011) for more details on these divisions.

Each entry in this module is of the following general type:

```
fun lemma_senseOffset_t : t ;
```

Keyword 'fun' declares each entry as a function of the type 't'. The function name is composed of lemma, sense offset and a type 't', where lemma and sense offset are same as in the Princeton WordNet, while 't' is one of the morphological types in GF resource grammars.

This abstract representation will serve as a pivot for all concrete representations, which are described next.

[2] This module has definitions of different morphological and syntactic categories in the GF resource grammar library

2.2 GF Concrete Lexicons

We build the concrete representations for different languages using the translations obtained from the Universal WordNet data and GF morphological paradigms (Détrez and Ranta, 2012; Bringert et al., 2011). The Universal WordNet translations are tagged with a sense offset from WordNet 3.0[3] and also with a confidence score. As, an example consider the following segment form the Universal WordNet data, showing German translations for the noun synset with offset '13810818' and lemma 'rest' (in the sense of 'remainder').

```
n13810818 Rest            1.052756
n13810818 Abbrand        0.95462
n13810818 Ruckstand     0.924376
```

Each entry is of the following general type.

```
posSenseOffset translation confidence-score
```

If we have more than one candidate translation for the same sense (as in the above case), we select the best one (i.e. with the maximum confidence score) and put it in the concrete grammar. Next, we give a small segment from the German concrete lexicon.

```
concrete LinkedDictGer of LinkedDictAbs = CatGer ** open
 ParadigmsGer, IrregGer,Prelude in  {
  lin consecutive_01624944_A =  mkA "aufeinanderfolgend" ;
  lin consequently_00061939_Adv =  mkAdv "infolgedessen" ;
  lin conservation_06171333_N =  mkN "Konservierung" ;
  lin conspire_00562077_V = mkV "anzetteln" ;
  lin sing_01362553_V2 = mkV2 (mkV "singen" ) ;
  ......
 }
```

The first line declares 'LinkedDictGer' to be the concrete representation of the previously defined abstract representation (note the keyword 'concrete' at the start of the line). Each entry in this representation is of the following general type:

```
lin lemma_senseOffset_t = paradigmName "translation" ;
```

Keyword 'lin' declares each entry to be a linearization of the corresponding function in the abstract representation. 'paradigmName' is one of the morphological paradigms defined in the 'ParadigmsGer' module. So in the above code, 'mkA', 'mkAdv', 'mkN', 'mkV' and 'mkV2' are the German morphological paradigms[4] for different lexical categories of 'adjective', 'adverb', 'noun', 'intransitive verb', and 'transitive verb' respectively. "translation" is the best possible translation obtained from the Universal WordNet. This translation is passed to a paradigm as a base word, which then builds a full-form inflection table. These tables are then used in the linearization phase of the translation system (see section 3)

Concrete lexicons for all other languages were developed using the same procedure. Table 1 gives some statistics about the coverage of these lexicons.

Language	Number of Entries	Language	Number of Entries
Abstract	91516	German	49439
French	38261	Finnish	27673
Swedish	23862	Hindi	16654
Bulgarian	12425		

Table 1: Lexicon Coverage Statistics

[3] However, in our concrete lexicons we match them to WordNet 1.7.1 for the reasons mentioned previously

[4] See (Bringert et al., 2011) for more details on these paradigms

3 System architecture

Figure 1 shows an architecture of the translation pipeline. The architecture is inter-lingual and uses the Resource Grammar Library (RGL) of Grammatical Framework (Ranta, 2011) as the syntax and semantics component, Penn Treebank data for parse-tree disambiguation and IMS(It Makes Sense)(Zhong and Ng, 2010) as a word sense disambiguation tool. Even though the syntax, semantics and parse-tree disambiguation are not the main topics of this paper, we give the full architecture to show where the work reported in this paper fits. Internal GF resources (e.g. resource grammars and dictionaries) are shown in rectangles while the external components (e.g. PennTreebank and IMS(Zhong and Ng, 2010): a wide coverage word sense disambiguation system for arbitrary text.) are shown in double-stroked rectangles.

With reference to Figure 1: The input is parsed using English resource grammar (EngRG) and a comprehensive English dictionary (DictEng). If the input is syntactically ambiguous the parser will return more than one parse-tree. These trees are disambiguated using a statistical model build from the PennTreebank data. The best tree is further processed using the input from the IMS to tag the lexical nodes with best sense identifiers. This tree is finally linearized to the target language using the target language resource grammar (TLRG) together with the target language lexicon (LinkedDict) discussed in section 2.

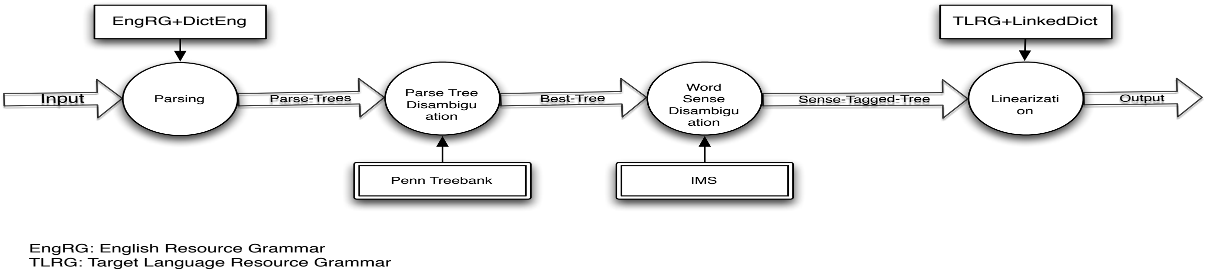
EngRG+DictEng
TLRG+LinkedDict
Input
Parsing
Parse-Trees
Parse Tree Disambigu ation
Best-Tree
Word Sense Disambigu ation
Sense-Tagged-Tree
Linearizati on
Output
Penn Treebank
IMS
EngRG: English Resource Grammar
TLRG: Target Language Resource Grammar

Figure 1: The translation pipeline.

4 Experimental Setup and Evaluation

Our experimental setup is as follows: We take some English text as source, and translate it to a target language (German and Hindi in these experiments) by passing it through the translation pipeline described in section 3. To show the usefulness of the lexicons described in section 2 and for comparison, we translate the same source twice: with and without word sense disambiguation.

For the first attempt, we used exactly the same translation pipeline as shown in Figure 1, except that to overcome the deficiencies of our existing parse-tree disambiguator, for some of the examples, we used trees directly from the PennTreebank, which are supposed to be correct. However, this should not damage the claims made in this paper which is about developing wide coverage interlingual translation lexicons and then using them for WSD in an interlingual translation pipeline.

For the second attempt, we plugged out the word sense disambiguation form the translation pipeline and used our old GF style lexicons (one target word per source word irrespective of its sense) in the linearization phase.

Finally, we compared both candidate translations to see if we have gained anything. We did both manual and automatic evaluations to confirm our findings.

For a set of 25 sentences for English-German pair we got marginal BLEU score improvements (from 0.3904 to 0.399 with 'old' and 'new' dictionaries). Manual inspection, however, was much more encouraging, and explained the reasons for very low improvements in the BLEU scores in some cases. The reason was that even if the word sense disambiguation, and hence, our new lexicon gives a better lexical choice, it will still be considered ‘wrong” by the evaluation tool if the gold-standard has a different choice. It was also observed that there were cases where the ‘old’ lexicon produced a much better translation than the ‘new’ one. The reasons for this are obvious. The word sense disambiguator has its own limitations and is known to make mistakes. Also, as explained in Section 5, the lexicon cannot be guaranteed to always give the right translation.

Next, we give a number of example sentence with comments[5] to show that how the new lexicons improved the quality of translations, and also give some examples where it worked the other way around.

4.1 German

1. **Source** He increases the board to seven

   Without WSD `er erhöht das Brett nach einigen sieben`

   With WSD `er vergrößert die Behörde nach einigen sieben`

   Comments `das Brett` is a wooden board (wrong); `erhöht` means “to raise”. while `vergrößert` means “increases the size”. Note the wrong preposition choice (“to” should be `zu` rather than `nach`). Also, an indefinite determiner (`einige`, some) has been wrongly added to the cardinal number is used as a noun phrase.

2. **Source** the index uses a base of 100 in 1,982

   Without WSD `das Verzeichnis verwendet eine Base nach einige 100 in einigen 1982`

   With WSD `der [index_11688271_N] nutzt einen Operationsbasis von einigen 100 in einigen 1982`

   Comments Note the untranslated word in the WSD version. `Base` means a chemical base, the wrong meaning here. `Operationsbasis` is not the best choice, but acceptable.

3. **Source** fear is the father of panic

   With WSD Angst ist der Papa von Angst

   Comment The traditional hilarious example, saying “fear is fear’s daddy”.

4.2 Hindi

To represent Hindi, we use an IPA style alphabet, with the usual values and conventions. Retroflexed sounds are written with a dot under the letter: ṭ, ḍ, and ṛ (a flap) are common, while ṇ and ṣ occur in Sanskritised Hindi (though many dialects pronounce them n and š). The palatalised spirant is shown š and aspirated stops are shown thus: k^{h}. A macron over a vowel denotes a long vowel, and ˜, nasalisation. In Hindi, e and o are always long, so the macron is dropped. Finally, we use ñ to mean the nasal homorganic with the following consonant.

Here are examples from our evaluation showing that the WSD system works well; the versions without WSD merely pick the first synonym in the lexicon.

1. **Source** Mr Baris is a lawyer in New York .

   Without WSD `Mr Baris New York mẽ kānūn kā pañḍit hæ`

   With WSD `Mr Baris New York mẽ vakīl hæ`

   Word order `Mr Baris New York in lawyer is`

   Comments `kānūn kā pañḍit` is “expert/teacher in law”, while `vakīl` means “lawyer”.

2. **Source** we don’t depend on pharmaceutical companies for our support

   Without WSD ham auṣad^{h}īya sahyōgī par hamāre b^{h}araṇ pōṣaṇ ke liye nahĩ nirte hæ̃.

[5] For the comments on German, we are indebted to Erzsebet Galgoczy and Wolfgang Ahrendt, our colleagues and German informants.

With WSD ham auṣad^{h}īya kañpanī par hamāre nirvāh vyay ke liye nahī̃ ūte hæ̃.

Word order We pharmaceutical companies on our subsistence expenditure for not ??? do

Comments sahyōgī means “company” in the sense of “colleagues”, nirvāh vyay means “subsistence expenditure” , while b^{h}araṇ pōṣaṇ means “weight bearing”. The penultimate word in both versions is nonsense, and the lexicons need to be debugged.

3. **Source** you may recall that a triangle is also a polygon

   Without WSD tum "recall may" ho ki ṭrāyengl "also" bahub^{h}uj hæ

   With WSD tum smaraṇ kar sakte ho ki trikoṇ b^{h}ī bahub^{h}uj hæ

   Word order You recall do can that triangle also polygon is

   Comments The version without WSD has several missing words. The WSD version of “recall” is not idiomatic, but understandable.

It should be noted that the coverage of the Hindi lexicon is lowest of all the lexicons given in Table 1. The result is that many sentences have missing words in the translations. Also, there is considerable interference with Urdu words (some stemming from the shared base grammar (Prasad and Shafqat, 2012)). Further, some mappings coming from the Universal WordNet data are in roman, as opposed to Devanagari (the usual script for Hindi, and what the grammar is based on), so these need to be transcribed. Finally, idiomatic phrases are a problem (“before the law” is likely to be rendered “(temporally) before the law” rather than “in the eyes of the law”).

5 The next steps

Since the Universal WordNet mappings are produced from parallel data by machine learning techniques, the translations are not always accurate and do not always make the best possible choice. This leaves a window for improvement in the quality of the reported lexicons. One way of improvement is the manual inspection/correction, not an easy task for a wide-coverage lexicon with around 100 thousand entries, but not impossible either. This would be a one-time task with a strong impact on the quality of the lexicon. Another way is to use manually built WordNets, such as the Finnish and Hindi WordNets. In our work, the availability of some of these resources was an issue, so we leave it for the future. Further, as mentioned in Section 4, the Hindi lexicon has some script-related issues which should be fixed in future.

When it comes to interlingua-based arbitrary machine translation, an important concern is the size of lexicons. We are aware of the fact that the size of our lexicons is not comparable to some of the other similar systems such as ATLAS-II (Fujitsu), where the size of lexicons is in millions. We have plan to extend the size of lexicons using some of the other publicly available resources (such as Hindi WordNet) and/or using parallel corpus. The development of bilingual lexicons form parallel corpus have been previously explored (Delpech et al., 2012; Qian et al., 2012), and the same ideas can be applied in our case.

6 Conclusion

We have shown how to use existing lexical resources such as WordNets to develop an interlingual translation lexicon in GF, and how to use it for the WSD task in an arbitrary text translation pipeline. The improvements in the translation quality (lexical), shown by examples in Section 4, are encouraging and motivate further work in this direction. However, it should be noted that there is still a lot of work to be done (especially in the open domain text parsing and parse-tree disambiguation phases of the translation pipeline) to bring the translation system to a competitive level. For the reasons noted in the introduction, we expect our techniques to be particularly useful for South Asian languages.

References

Angelov, K. (2011). *The Mechanics of the Grammatical Framework.* PhD thesis, Chalmers University Of Technology. ISBN 978-91-7385-605-8.

Angelov, K. and Enache, R. (2010). Typeful Ontologies with Direct Multilingual Verbalization. In Fuchs, N. and Rosner, M., editors, *CNL 2010, Controlled Natural Language.*

Bringert, B., Hallgren, T., and Ranta., A. (2011). GF resource grammar library synopsis. www.grammaticalframework.org/lib/doc/synopsis.html.

Curry, H. B. (1961). Some logical aspects of grammatical structure. In Jakobson, R., editor, *Structure of Language and its Mathematical Aspects: Proceedings of the Twelfth Symposium in Applied Mathematics*, pages 56–68. American Mathematical Society.

de Melo, G. and Weikum, G. (2009). Towards a Universal Wordnet by learning from combined evidence. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pages 513–522, New York, NY, USA. ACM.

Delpech, E., Daille, B., Morin, E., and Lemaire, C. (2012). Extraction of domain-specific bilingual lexicon from comparable corpora: Compositional translation and ranking. In *Proceedings of COLING 2012*, pages 745–762, Mumbai, India. The COLING 2012 Organizing Committee.

Détrez, G. and Ranta, A. (2012). Smart paradigms and the predictability and complexity of inflectional morphology. In *EACL*, pages 645–653.

Dymetman, M., Lux, V., and Ranta, A. (2000). XML and multilingual document authoring: Convergent trends. In *Proc. Computational Linguistics COLING, Saarbrücken, Germany*, pages 243–249. International Committee on Computational Linguistics.

Enache, R., España-Bonet, C., Ranta, A., and Márquez, L. (2012). A hybrid system for patent translation. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT12), Trento, Italy.*

Hindi-WordNet (2012). *Hindi Wordnet. 2012. Universal Word—Hindi Lexicon.* http://www.cfilt.iitb.ac.in.

Lindén, K. and Carlson., L. (2010). Finnwordnet—wordnet på finska via översättning. *LexicoNordica—Nordic Journal of Lexicography*, 17:119–140.

Ljunglöf, P. (2004). *The Expressivity and Complexity of Grammatical Framework.* PhD thesis, Dept. of Computing Science, Chalmers University of Technology and Gothenburg University. `http://www.cs.chalmers.se/~peb/pubs/p04-PhD-thesis.pdf`.

Lyons, J. (1968). Introduction to theoretical linguistics. *Cambridge: Cambridge University Press.*

Martin-Löf, P. (1982). Constructive mathematics and computer programming. In Cohen, Los, Pfeiffer, and Podewski, editors, *Logic, Methodology and Philosophy of Science VI*, pages 153–175. North-Holland, Amsterdam.

Miller, G. A. (1995). Wordnet: A lexical database for English. *Communications of the ACM*, 38:39–41.

Montague, R. (1974). *Formal Philosophy.* Yale University Press, New Haven. Collected papers edited by Richmond Thomason.

Prasad, K. V. S. and Shafqat, M. V. (2012). Computational evidence that Hindi and Urdu share a grammar but not the lexicon. In *The 3rd Workshop on South and Southeast Asian NLP, COLING.*

Qian, L., Wang, H., Zhou, G., and Zhu, Q. (2012). Bilingual lexicon construction from comparable corpora via dependency mapping. In *Proceedings of COLING 2012*, pages 2275–2290, Mumbai, India. The COLING 2012 Organizing Committee.

Ranta, A. (2004). Grammatical Framework: A Type-Theoretical Grammar Formalism. *The Journal of Functional Programming*, 14(2):145–189. `http://www.cse.chalmers.se/~aarne/articles/gf-jfp.pdf`.

Ranta, A. (2011). *Grammatical Framework: Programming with Multilingual Grammars.* CSLI Publications, Stanford. ISBN-10: 1-57586-626-9 (Paper), 1-57586-627-7 (Cloth).

Ranta, A. and Angelov, K. (2010). Implementing Controlled Languages in GF. In *Proceedings of CNL-2009, Athens*, volume 5972 of *LNCS*, pages 82–101.

Ranta, A., Détrez, G., and Enache, R. (2012). Controlled language for everyday use: the MOLTO phrasebook. In *CNL 2012: Controlled Natural Language*, volume 7175 of *LNCS/LNAI*.

Rosetta, M. T. (1994). *Compositional Translation*. Kluwer, Dordrecht.

Seki, H., Matsumura, T., Fujii, M., and Kasami, T. (1991). On multiple context-free grammars. *Theoretical Computer Science*, 88:191–229.

Shafqat, M., Humayoun, M., and Aarne, R. (2011). An open source Punjabi resource grammar. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 70–76, Hissar, Bulgaria. RANLP 2011 Organising Committee. http://aclweb.org/anthology/R11-1010.

Stallman, R. (2001). *Using and Porting the GNU Compiler Collection*. Free Software Foundation.

Zhong, Z. and Ng, H. T. (2010). It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, pages 78–83, Uppsala, Sweden. Association for Computational Linguistics. http://www.aclweb.org/anthology/P10-4014.