UNIVERSITÀ DI PISA

DIPARTIMENTO DI
FILOLOGIA, LETTERATURA E LINGUISTICA

CORSO DI LAUREA MAGISTRALE IN
INFORMATICA UMANISTICA

TESI DI LAUREA

Cross-Lingual Word Sense Annotation
with Multilingual WordNets

CANDIDATA
Giulia Bonansinga

RELATORI
Prof. Alessandro Lenci
Prof. Francis Charles Bond

CONTRORELATORE
Dott. Felice Dell'Orletta

ANNO ACCADEMICO 2017/2018

Abstract

Cross-lingual approaches are exploited to enrich existing parallel corpora with semantic annotation in an inexpensive fashion. Human-checked annotations, though extremely beneficial to make substantial progress in Word Sense Disambiguation (WSD), are very time-consuming to produce and alternative options should be explored.

We first compare two such approaches that can be applied to any multilingual parallel corpus, as long as large inter-linked sense inventories exist for all the languages involved and word alignments are provided. If not complete, at least partial disambiguation can be achieved by exploiting both the similarities and differences among the languages involved. Secondly, we attempt to disambiguate a multilingual parallel corpus, derived from SemCor and its sibling projects (Bentivogli and Pianta 2005; Bond et al. 2012; Landes et al. 1998; Lupu et al. 2005), by means of Multilingual Sense Intersection (MSI).

Unlike sense projection, MSI can be applied to most existing multilingual parallel corpora, because it does not require the availability of sense annotation for any text in the corpus. MSI, though more error-prone, can boost coverage of the annotation for multilingual parallel corpora, as long as there are sense inventories of adequate size linked to each other for the target languages.

The availability of sense-annotated corpora is crucial for training Supervised WSD systems and make further progress in many other Natural Language Processing (NLP) tasks. We release the tools to perform MSI and the result of its application on a subset of the SemCor projects.

Acknowledgments

This research was supported in part by the Erasmus Mundus Action 2 program MULTI of the European Union (grant agreement number 2010-5094-7) and the MOE Tier 2 grant *That's what you meant: a Rich Representation for Manipulation of Meaning* (MOE ARC41/13).

I dedicate this work to my support system - you know who you are.

Contents

List of Figures

List of Tables

Glossary

AAR Average Ambiguity Reduction. 108

CL-WSD Cross-lingual Word Sense Disambiguation. 5, 12, 22, 23, 27

DTD Document Type Definition. 118

ILI InterLingual Index. 31, 74

IR Information Retrieval. 21

IWN ItalWordNet. 73

JSC Japanese SemCor. 59

JWN Japanese WordNet. 60, 76

MFS Most Frequent Sense. 18, 20, 105

MPC Multilingual Parallel Corpus. 63, 64, 125

MSC MultiSemCor. 33, 37, 39, 50, 59

MSI multilingual sense intersection. x, xii, 23, 41, 43, 97, 99, 108, 113, 129

MT Machine Translation. 4, 21, 25, 38

MWN MultiWordNet. 35, 40, 72

NLP Natural Language Processing. ii, 21

OMW Open Multilingual WordNet. 78, 101

PWN Princeton WordNet. 48

RSC Romanian SemCor. 57, 131

SC SemCor. 32, 34–36

SFS sense frequency statistics. 93, 101, 129

SP sense projection. 33, 37–39

WN WordNet. 6, 26, 29, 35, 66

WN 3.0 WordNet 3.0. 63

WSD Word Sense Disambiguation. 3–5, 7, 12, 16, 17, 21, 22, 26, 27, 38, 103

Chapter 1

Introduction

1.1 Processing language

The ability to understand language is considered as a fundamental test (Turing 1950) in the Artificial Intelligence (AI) domain. In order to be able to claim that a machine understands language, it is crucial for the machine to have means to solve natural language ambiguities and discern the intended sense in context of any polysemous word.

This comes very naturally to humans in their daily tasks, when they can rely on extra-linguistic context and word knowledge. Yet, from a computational point of view, Word Sense Disambiguation (WSD) is still unsolved and commonly regarded as an AI-complete problem.

The first appearance of WSD as a computational problem took place in the domain of Machine Translation (MT), very early on. In his seminal memorandum on the topic, Weaver (1949) suggested that the problem of ambiguity and the consequent issue of choosing the most appropriate translation could be aided by the observation of context. And WSD is, indeed, the problem of guessing the intended sense of a word in a given context and thus constitutes a requirement for MT.

There is no agreement upon what represents a word meaning. A lexicographer holds the responsibility to give a full account of the words of a language, listing out all their meanings. Provided a certain theoretical standpoint to the mental lexicon, the approach usually followed comes down to see meanings as abstractions of pattern of use (Kilgarriff 2006). This is especially reasonable in an application-oriented view.

However, dictionaries vary greatly in size and detail, and the smaller ones often group together fine-grained senses and generalize them under a coarser meaning. What can really be expected to be in *any* dictionary for a given entry is the senses that are 'sufficiently frequent and insufficiently predictable', as stated in that Kilgarriff's paper revealingly titled as *I don't believe in word senses* (1997).

One more practical distinction to get a grasp of what constitutes a word sense was suggested by Resnik and Yarowsky (1997), who prompted that, for the purposes of WSD, we could rely on the sense distinctions that are lexicalized cross-linguistically. In fact, one would expect that if other languages has more lexicalizations for a certain word, there must be conceptual grounds for it. This especially makes sense in the several Cross-lingual Word Sense Disambiguation (CL-WSD) approaches that will be presented in Chapter 2.

The goal of *computational* WSD is to determine automatically which sense of a word is activated in a particular context (Navigli 2009). In this formulation, WSD is basically a classification task for each polysemous word, provided a lexical knowledge base in which words are associated with discrete sets of senses (Agirre and Edmonds 2007). Such a resource is challenging to compose, thus posing an additional difficulty to the problem of agreeing on a configuration where the lists of senses need to be as shared and exhaustive as possible.

1.2 A brief history of Word Sense Disambiguation

Ambiguities in language started to be investigated back in the late 1940s, but it was not until the 1980s that the way of doing research in WSD underwent a deep revolution, coincidentally with the advent of large-scale lexical resources.

Lesk's approach (1986) belongs to this turning point. In its original design, the algorithm tries to detect the correct word sense by using a Machine-Readable Dictionary (MRD) and looking at the overlap of the candidate senses' definitions with the context. This simple idea has inspired much following research, whose first aim has been to overcome its main Achilles' heel, i.e. the sensitiveness to the exact wording of the glosses.

The following decade continued in the same vein and the WSD task really started to be tackled gaining from a whole new range of approaches; the large employment of statistical methods belongs to this phase. However, what really boosted research in WSD was the launch of WordNet (WN), a large lexical database hierarchically organized into word senses (Miller et al. 1990, see Section 3.2.1). WN immediately became the most used resource for dictionary-based approaches.[1]

Senseval evaluation campaigns (see Section 1.3) began shortly after and marked a fundamental progress in the way the research in the field was being done, proposing a standardized way of testing and evaluating competing systems. Ide and Véronis (1998) offer a very good recap of milestones, approaches and results up to 1998.

1.2.1 Approaches to WSD

In a monolingual setting, any clues about the intended sense of a word have to be contextual in nature. If a word w is used in two equivalent contexts, then we say w represents the same sense in both of them. There have been many attempts to formalize the notion of context and make it quantifiable and there are just as many models as the variability introduced by the features considered. Tufiş et al. (2004) listed some of the most common. When taking into account the context length, for instance, the context window could be limited to the immediately preceding/following words or to the full sentence.

[1] Unfortunately, such hand-crafted inventories pose a risk in that they often lack complete coverage of sense distinctions; see also chapter 3 and 5 for the issues encountered using WordNet.

The context content allows even more variability; for instance, one could take into account only the words that are at a particular distance or in a certain grammatical relation to the target word. Additional weights can then be considered in the parameter tuning; for instance, a distance-based measure could be included to weigh words closer to the target word more.

However, there are more than local features; the whole sentence, paragraph or even document can become (topical) features themselves, typically in a bag-of-words form.

Context formalization represents a more complex issue. It would be helpful to be able to compare different contexts; in order to do, contexts need to undergo some sort of formalization. For instance, a context for a specific target word could be reduced to the parts of speech of the words in its context window or to their logical functions, thus including syntactic features. If previous words in the text have already been assigned a sense tag, even this information can be used as a semantic feature. The possibilities really are innumerable.

Vectors are a typical way to represent many features. Navigli (2009) observes that flat vectors are more apt for supervised approaches, while more structured representations can be fully exploited, in their range of lexical and semantic relations, by unsupervised and knowledge-based methods.

Features aside, the approaches to WSD vary according to the methodology and the resources assumed, if any.

Knowledge-based approaches

Also referred to as dictionary-based methods, they solely rely on thesauri, dictionaries or lexical knowledge bases such as WordNet (see Section 3.2), from which they draw human-made sense distinctions, thus not making use of corpus-derived evidence. Lesk's algorithm falls in this category. Graph-based methods are a recent addition to knowledge-based methods, in which nodes are senses and edges represent semantic relations between nodes; see for instance Navigli and Lapata (2007) and Sinha and Mihalcea (2007).

While it comes naturally to think of structured resources as the ones just named to the aid of knowledge-based methods, those are not the only ones. There are other unstructured sources of information that are often employed to ease the disambiguation process, such as corpora (obviously), collocation resources, word frequency lists, sense frequency lists and stoplists (Navigli 2009).

Supervised approaches

Supervised methods learn from annotated corpora or bootstrap from seed data (being so semi-supervised). In such approaches, given a set of words to disambiguate, a corpus is collected where each occurrence of any word in this set is manually annotated with its correct sense. The corpus built with respect to these requirements is then used to train a learning algorithm, that will learn a model that will be used with unseen occurrences of the target word set.

Provided a sufficient number of sense-tagged examples, supervised methods are going to give the best results. Nevertheless, the requirement is not as feasible to meet as one would think: Palmer et al. (2006) estimate that there are 75,000 polysemous WordNet sense tags for English, so a training corpus with several sense-tagged occurrences for each sense would be required. Such an enormous quantity of sense-annotated data has not been produced and will not be available any time soon, as the task is extremely time-consuming and hard even for trained annotators. This problem, known as **knowledge acquisition bottleneck**, offers motivation for investigating alternative methods to produce sense-annotation; see chapters 2 and 3.

Mihalcea and Moldovan (1999) proposed early on a valid method to acquire sense-tagged examples in order to bypass the knowledge acquisition bottleneck, which basically makes use of Princeton WordNet and information that can be found on the Internet through queries employing logical operators. Specifically, for each sense the method retrieves a monosemous synonym in the synset, if any, and otherwise exploits the information in the gloss to produce a number of sense-specific example sentences that are validated through web search. This procedure was applied to 20 polysemous words of various parts of speech that altogether had 120 word senses. A maximum number of 10 examples was retained for each word. Of the 80,741 example sentences acquired, 1,080 were manually checked, giving an estimated accuracy of 91%.

There are at least two common techniques to relieve the knowledge acquisition bottleneck, besides of course the manual annotation of new corpora. The first consists in bootstrapping and active learning (Navigli 2009). Starting from a small amount of annotated data and large amounts of unannotated data, a set of classifiers is used iteratively to bootstrap annotation for the latter, until a threshold is reached; a famous example is work by Li and Li (2004) (see Section 2.1 for a description). The second technique is based on exploiting cross-lingual information to cheaply annotate large amounts of text, which is the domain of investigation of the present work.

Unsupervised approaches

Unsupervised methods do not rely on any external knowledge source, so word senses are usually induced by clustering word occurrences from unlabeled training data. The assumption behind is that several instances of a word in a specific sense will have similar contexts. Two criteria can be followed: the first is based on distributional similarity; for instance, words can be clustered together with respect to the number of overlapping context words; the second consists in inferring sense distinctions from translation pairs in aligned parallel corpora, i.e. on *translational equivalence* (Lefever and Hoste 2014). The classes found in this fashion are then used to classify new data.

The main problem with fully, pure unsupervised methods is that they really identify sense clusters, rather than assigning sense labels from a shared sense inventory (which is not employed), so in the end the data thus annotated are not 'sense-annotated data' in the usual sense. However, as Navigli (2009) points out, identifying sense clusters certainly is a WSD subproblem and very much related to it.

To each problem its own solution

As long as performance is concerned, supervised methods perform best, but they come with a considerable problem in their need for large training corpora. Ng (1997) estimated that a corpus with no less of 3.2 million sense-tagged words would be needed to train a supervised system with very wide coverage.

Due to this dependence on sufficiently large training data, a supervised system is not always the most reasonable choice, especially for languages with fewer resources. Knowledge-lean methods, on the other hand, are likely to be preferred because the requirements they rely on are easier to meet; in fact, lexical resources such as WordNet are constantly enriched and improved (and getting more and more multilingual, see Section 3.2), so their performance improves accordingly (Navigli 2009).

Besides, just as supervised systems often have to back-off to knowledge-based strategies to break through the impasse of insufficient training data, the same occurs with knowledge-based methods, that often recur to first sense heuristics and the like.

1.2.2 Getting multilingual

As mentioned, parallel corpora have proved to be a precious ally early on in WSD research, and they have also been employed in combinations with knowledge-based methods.

In a multilingual setting, the focus on context remains but changes its nature, the context of a word becoming reciprocal translations in different languages. The intuition behind this change of perspective says that translation preserves meaning, and thus parallel text selects the correct context among the various possible translations. As pointed out by Tufiş and Ion (2004) and Tufiş et al. (2004), the employment of parallel corpora cannot help in resolving all ambiguities, because often the polysemy is preserved in the other language. Also, one cannot assume that different lexicalizations of a certain source word in the target language always occur for valid conceptual reasons.

This thesis especially focuses on **CL-WSD** and the contribution brought by different languages, as discussed in section 1.5. Tufiș and Ion (2003) claim that a multilingual approach to WSD can offer far more precise insight into word meaning than traditional monolingual methods.

In their 2012 essay, Bandyopadhyay give a comprehensive overview of the most influential and groundbreaking approaches explored till then; see Fig. 1.1 and 1.2.

As mentioned above and endorsed by Agirre and Stevenson (2006), the va-

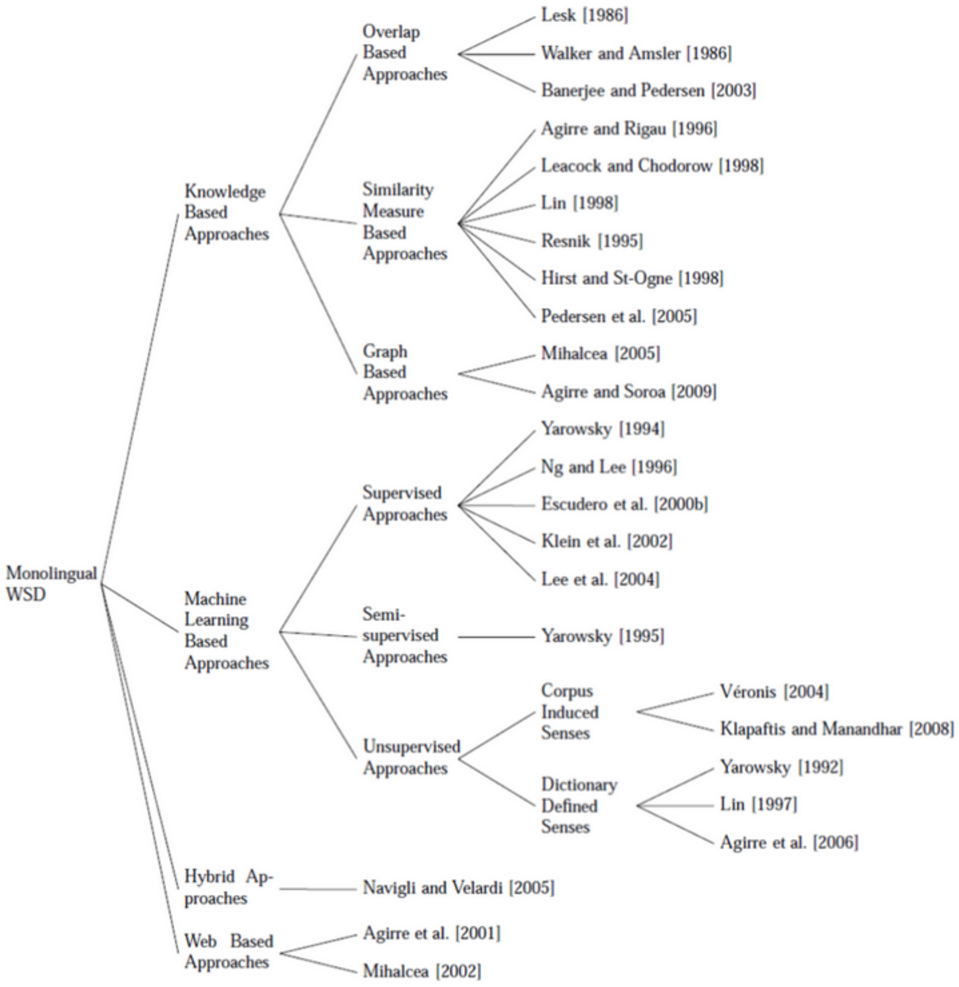


Figure 1.1: A taxonomy of monolingual approaches to WSD. From Bandyopadhyay (2012, p. 25)

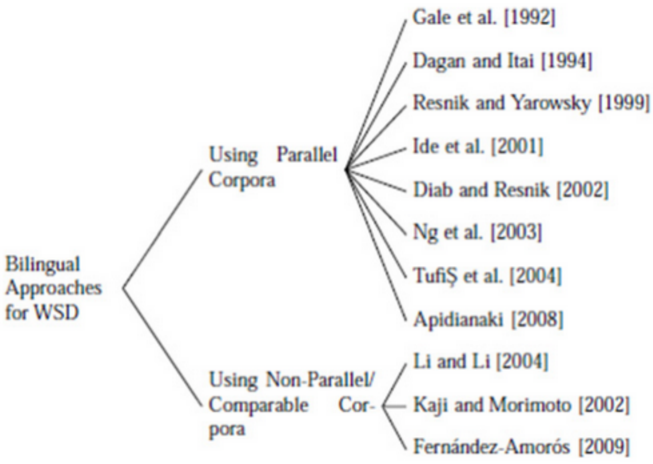


Figure 1.2: A taxonomy of bilingual approaches to WSD. From Bandyopadhyay (2012, p. 32)

riety in WSD approaches is mainly due to the abundance of features that can be employed in models to enhance the disambiguation process; anything from part-of-speech, lemma, a context-window to complex argument structure constructions and semantic domains may give a contribution, thus the enormous range of algorithms proposed in the literature.

1.2.3 Words are not all the same

When thinking about meaning, the first distinction that should be made is between vagueness and ambiguity. While a vague word has one meaning that lacks 'perfectly well-defined boundaries' and is general enough to be used to pertain to many different things, ambiguity takes place whenever a linguistic expression that 'can have more than one distinct denotation' (Murphy 2010; Wasow et al. 2005). It follows that vagueness is more of a property that a sense of a lexeme has, while ambiguity concerns the relation between lexeme(s) and senses.

Among the ambiguous words, one can name two further subclasses.

Homonyms just happen to share the same written form, but have unrelated meanings and often different etymology. On the other hand, polysemous words root back to the same lexeme and have two or more distinguishable, but related senses (Murphy 2010, p84).

Distinguishing between vagueness and polysemy can be challenging even for native speakers. Murphy proposes the fitting example of *friend*, which is vague with respect to the gender; i.e., it is not the case of two distinct senses for *female friend* and *male friend*. This can easily be proved by a few tests: first, both alleged senses of *friend* can go under the same definition (*definition test*); secondly, using two instances of *friend* in the same sentence sounds somehow wrong (*I have friends and friends*), because the two instances do not refer to separate senses and do not contrast with each other (*contrast test*). Finally, if two senses of the same lexeme are intended in the same sentence, one should experience a strange feeling that goes under the name of *zeugma effect*. If this does not occur, then we are in presence of a single sense, although vague.

Informally, the term ambiguity is often used in opposition to polysemy to refer to a property of text, whenever it is not clear what message is conveyed. Polysemy, on the other hand, is an intrinsic property that words show in isolation from text, and that can be solved looking at the context for clarification (Agirre and Edmonds 2007, p8).

WSD tends to be considered a solved problem for homographs, which are words that happen to have the same spelling but completely unrelated meanings. The sense distinctions in homographs are coarse-grained, while in polysemous words senses are related or are different aspects of the same denotatum, so they can be distinguished at a finer level.

The well-known Zipfian law (Zipf 1949) that accounts for the skewed distribution of many physical and social phenomena also applies to the distribution of word senses. The more frequent the word, the more senses it will have (and vice versa) in a power-law relationship. Ng and Lee (1996) calculated that the 121 most frequent English nouns have on average 7.8 meanings each in the first version of WordNet.

1.3 Evaluation campaigns for WSD

Senseval[2] is an international competition for the evaluation of WSD systems, started in 1998 and then regularly held every three years till a new course started and its name changed in Semeval, in 2007.

Resnik and Yarowsky (1999) initiated work on Senseval by giving their insight on evaluation criteria; in fact, even basic building blocks such as evaluation metrics

[2] www.senseval.org

had to be agreed upon in preparation for shared tasks (Palmer et al. 2006).

The first online evaluation exercise proposed the disambiguation of a *lexical sample*. This task, also commonly referred to as targeted WSD, focuses on the full disambiguation of a limited set of highly polysemous words. Participants were provided with training and test data and an agreed-upon metric for evaluation, so that their systems could be easily compared.

Lexical sample WSD suits better supervised systems, that can be trained using an adequate number of hand-labeled examples. Knowledge-based systems, on the other hand, will typically need some sort of heuristics or information other than the plain list of possible senses. The already mentioned Lesk's approach, for instance, gains more information by employing both the context words and the sense glosses.

The motivation for evaluation campaigns arose from the need of coherently comparing and evaluating different systems for WSD and being able to measure progress over time. With Senseval and the competitions that followed up, manually annotated training corpora and benchmark test data were produced and made available for further research, giving a clear direction to follow and agreed metrics to measure progress over the same data.

1.3.1 From Senseval to Semeval

Senseval-1 (Kilgarriff and Palmer 2000; Kilgarriff 1998) consisted of a lexical-sample task for English, French and Italian. 23 research groups, to a total of 25 systems, entered the first competition, challenging themselves with a test set of 35 target words in 8,400 instances. The best systems achieved 74–78% accuracy, while the Most Frequent Sense (MFS) baseline scored 57%.

Since its beginning, the campaign has kept opening up to several kinds of tasks and different languages. Senseval-2 (Edmonds and Cotton 2001) introduced an *all-words* task along with the lexical sample task. In the all-words WSD, all open-class words in a running text are to be disambiguated. Supervised systems usually plod on this task because the large amount of training examples required, while knowledge-based approaches that lean on wide-coverage resources naturally have better chances.

Participation had increased as well: 93 systems joined the competition, and WordNet 1.7 was being introduced as the shared common sense inventory for English. Its fine granularity is probably the reason why the performance actually dropped from the previous campaign. Nevertheless, the shared effort was towards WordNet because of its popularity and availability, and supervised systems yet achieved, all in all, better performance than knowledge-lean systems.

The task was harder for unsupervised systems, understandably, which performed well below the first sense baseline (48% accuracy), scoring 40%.

What was figured back then and still holds nowadays is that knowledge-based and supervised machine learning systems do obtain the best results, but are hampered by the dependence on wide-coverage sense inventories and sense-tagged training corpora. As a result, the need for unsupervised techniques remains as strong as always (Ion and Tufiş 2009).

Mihalcea and Edmonds (2004) organized Senseval-3, including seven languages and new tasks such as as gloss disambiguation, multilingual annotation, semantic role labeling, subcategorization information acquisition and logic forms. This was the campaign when the lexical sample task started to be regarded to as less interesting; among the best systems, the peak reached was 72.9%, close to human levels and, at the same time, seemingly unable to outdo the plateau that had been reached. Moreover, the community started to feel that the disambiguation of a single target word (although particularly rich semantically) was not the best place to spend energies on.

Three years later, the fourth edition started a new course with Semeval-2007 (Agirre et al. 2007), to make clear that the domain of the evaluation exercises had been extended to tasks of semantic analysis other than WSD. This edition saw the insertion of a cross-lingual task, since recurring (see also Section 2.2.3 for cross-lingual WSD work tested through Semeval).

Semeval maintained the three-year break for another edition, Semeval-2010 (Erk and Strapparava 2010), then the following was split in Semeval-2012 (Agirre et al. 2012) and Semeval-2013 (Manandhar and Yuret 2013) depending on the tasks considered, and from then on the competition has been held annually till nowadays (Nakov and Zesch 2014; Nakov et al. 2015), even though not all tasks are run every year.

1.3.2 How far did we go?

Navigli (2009) gives a full account of best systems and results, but above all he proposes a reflection on the feasibility and the shortcomings of evaluation exercises.

As a first issue, the dictionaries employed as sense inventories have changed often, making a comparison between the first campaigns especially difficult. In Semeval-2007 the maximum score achieved with the fine-grained all-words task was 65–70%, not less than ten percentage points with respect to other works using coarse-grained senses. This led to the organization of coarse-grained tasks in the same edition, reaching 88.7% and and 83% accuracy respectively in the lexical sample and the all-words tasks.

Navigli also points out that it is hard to properly weigh the contribution of different techniques in the systems proposed, as they interact with each other and cannot really be evaluated on their own. Additionally, the MFS baseline was considered very hard to beat in all-words systems back to ten years ago, and it is still today, affecting closely also the present work.

1.4 WSD for applications

WSD is a crucial module for several applications in NLP, among which parsing, MT, Information Retrieval (IR) come immediately to mind.

WSD contributes to MT in that it helps choose the correct translation for polysemous words, especially in terms of idiomaticity and fluency. A well-known unsupervised approach by Brown et al. (1991) incorporated a WSD component in the analysis phase of a traditional statistical machine translation system. After word-aligning the parallel corpus, the most appropriate translation for a target word is chosen based on the most telling feature in a predefined set of contextual features. As a result, the different senses of words in the target language are labeled, bringing down the overall error rate by thirteen percent.

In IR the other query terms usually give sufficient clues for the correct interpretation of an ambiguous word (e.g., ”interest rates” already rules out documents conveying the *hobby* interpretation). Sanderson has observed in more than one occasion that WSD cannot really boost IR if long queries are submitted, and also that the real challenge lies in uncommon ambiguous words, rather than very frequent ones, as the document context comes to the aid of the latter (Sanderson 1994). Schutze and Pedersen (1995) presented encouraging results on the contribution of WSD to IR: provided a WSD system with accuracy greater than 90%, then the IR system's performance gains an additional 4.3% in precision.

Other tasks naturally rely on a particularly accurate text analysis and necessitate a built-in WSD module. This is the case of Information extraction (IE) and text mining, as well as more specific subproblems such as Named-Entity Recognition (NER) or acronym expansion (Navigli 2009). Task 8 in Semeval-2007 broached the IE domain by requiring participant systems to select the right metonymy for target named entities (Markert and Nissim 2007).

There is a longtime debate on whether research in WSD should be carried out with an application in mind - that is *in vivo*, as an integrated component - or on its own, *in vitro*, using benchmarks designed for the purpose. While the advancement of research in WSD becomes immediately apparent when investigated within a specific application, it is easier to compare and evaluate different strategies independently from other tasks.

The other fundamental issue is understanding how significant and how detailed the contribution of a WSD module should be in these applications. Ide and Wilks (2006) strongly suggest that disambiguation at the homograph-level is adequate for most NLP applications; additionally, at this level of granularity WSD algorithms perform well and are very robust.

1.5 Scope of this work

CL-WSD differs from WSD in that it makes use of parallel corpora and exploits similarities and differences in languages to disambiguate one another. This formulation of the WSD problem is thus very specific and only applies to certain target texts. Yet, it allows to inexpensively tag large amounts of data with semantic annotation.

To clarify how this can be carried out, let us examine the task of disambiguating all the content words, in each language, in a multilingual parallel corpus. If we are able to compare each word in running text with its aligned translations, then we have access to more semantic information and we can make more educated guesses of the meaning intended in context by *all* the aligned translations. In fact, looking at translation pairs we have more chances to pinpoint the actual sense activated by a polysemous word. This simple, yet powerful, analysis of **cross-lingual lexicalization** can help to reduce or even solve the ambiguities and thus the human effort in annotating a whole text from scratch.

If such an approach proves itself to be an alternative precise enough to replace manual annotation, then we will be one step closer to easing the knowledge acquisition bottleneck.

The contents of this thesis are organized as follows. Chapter 2 reviews the main approaches to CL-WSD and the most influential works that made use of parallel corpora, with special attention to those focused on sense annotation. In chapter 3 the resources employed in this thesis are described in detail, followed by a discussion on the preprocessing steps and the requirements that need to be met in order to carry out the MSI procedure described in section 2.2.2.

The core of the approach explored in this thesis is described in chapter 4, with a discussion of the implementation details of MSI. This methodology is applied to a multilingual corpus of English, Italian, Romanian and Japanese parallel texts and followingly evaluated. The results achieved are then compared to those obtainable using coarse-grained senses.

Chapter 5 proposes a generalized procedure to make the work of this thesis reusable and available to the public for the task of enriching any given parallel corpus with sense annotation.

Finally, in chapter 6 the feasibility of the approach presented is discussed in light of the results given in chapter 4 and the interchange format presented in chapter 5.

Chapter 2

Related Work

2.1 The advent of parallel corpora in WSD

Parallel corpora made their first appearance as a source of training and test data for WSD in the early 1990s, even though the initial motivation had originated in MT, in the effort of finding a means to give the most reasonable translation in context for ambiguous words. Besides the already mentioned application in Brown et al. (1991), one other well-known contribution comes to mind: Dagan et al. (1991) showed that lexical ambiguities in the source language can be solved by looking up the candidate lexical relations in corpora in the target language and then relying on statistical data to select the most idiomatic translation, thus only exploiting a bilingual dictionary and a monolingual target language corpus.

Shortly afterwards, parallel corpora started to being exploited purposely for WSD, causing a turning point in the way research was carried out. In fact, translation information was finally providing a source of training and testing data other than hand-labeled examples, and far easier to acquire.

Gale et al. (1992c) exploited a supervised learning algorithm for WSD for which the training data was collected from parallel corpora, by aligning the reciprocal translations and feeding them to a classifier. In this way, real use sense examples can be extracted from parallel corpora for each sense of any polysemous word, with different translation choices in the target language being clues of distinct meanings involved. In the testing phase, the contexts of unseen examples can be compared against the training sense examples in order to assign the correct sense. While effective otherwise, this approach falls short whenever the ambiguity holds across languages, i.e. in cases of *parallel polysemy*, which is frequent for related language pairs.

Cross-lingual lexicalization has also been successfully exploited to validate existing sense inventories, such as WN; see for instance Ide (2000) and Resnik and Yarowsky (1999). Ide et al. (2002) made one step further by testing the sense distinctions found in an automatic way in a real WSD task, that is, without relying on external resources for validation. Specifically, they employed a parallel corpus built upon aligned versions in seven languages of George Orwell's *1984* and extracted *lexical translation equivalents*[1] from it.

[1] This expression is commonly used to refer to a symmetric relation between parts of a text in different languages that are reciprocal translations of one another.

In this approach, the translation equivalents were clustered in equivalence classes to automatically discern sense distinctions, which were later used in two experiments and proved to be as functional as those made by human annotators. This particular item of research showed very early on how sense disambiguation achieved in one language of a parallel corpus can be beneficial to the disambiguation of any aligned translation in another language, fully embracing the idea of CL-WSD for which one language helps disambiguate another.

Following this lead, Ng et al. (2003) exploited an English-Chinese parallel corpus (consisting of 31.7 million words in the English side and 55.2 in the Chinese side) to acquire sense-tagged training data to be fed to a WSD classifier. The training data obtained in this fashion proved to be sufficiently abundant and precise when used for disambiguating English nouns in unseen contexts extracted from the Senseval-2 English lexical sample task, with a 14% difference in accuracy (Carpuat and Wu 2007; Edmonds and Cotton 2001). In following work, Chan and Ng (2005) attempted to gather sufficient training data to disambiguate the nouns of Senseval-2 English all-words task. With the training examples extracted from 680 MB of English-Chinese parallel text, they performed as well as the best systems; moreover, even in its simplest configuration, this approach significantly outperformed the baseline of choosing the first sense listed in WordNet (61.1% accuracy versus 69.6%, with a further peak of 72.7% in a more refined setting). Unlike their previous experiment, where senses were lumped together if they were translated the same in Chinese, in this work a fine-grained disambiguation setting was employed. It is noteworthy that the training examples automatically acquired (a maximum of 500 for each noun) were precise enough to outperform the results obtained using the manually produced (but far fewer) sense examples extracted from SemCor.

Li and Li (2004) turned sense disambiguation in a bilingual setting in a word translation disambiguation problem, i.e. in the problem of selecting the correct lexical choice. They implemented a machine learning system that repeatedly builds and improves classifiers in both of the two languages using only little classified data in one language (optionally also in the other) and large amounts of unclassified data, that gradually are classified and contribute to boost the performance of the classifiers. Applied on English-Chinese bitext, this method, named *bilingual bootstrapping*, uses Chinese words in place of the English sense labels to tag the English words. This bootstrapping procedure took inspiration from work by Yarowsky (1995) performed in a monolingual setting.

Diab's unsupervised method - Many others have gone down the same road and devised unsupervised methods that exploit parallel corpora to avoid being dependent on hand-tagged data. Diab (2003) devised an unsupervised approach for the automatic sense annotation of parallel corpora, called Sense Assignment Leveraging Alignment And Multilinguality (SALAAM). In this framework, word meaning is 'quantifiable as much as it is uniquely translated in some language or set of languages' and thus can be approximated in a cross-linguistic view.

Given a bitext, source words that translate to the same target word are clustered together. For each cluster, the similarity among the different senses of the words in the source groups is measured using the thesaurus-based metric proposed by Resnik (1995) to disambiguate noun groupings with respect to the WN hierarchy. To determine how similar two words are, the metric basically exploits the *information content* of the most specific hypernym concept in common.

For each cluster consisting of at least two words, Resnik's measure is used to select the closest sense tag. The word-sense matches obtained in this fashion are first propagated onto the source corpus, and then from there to the corresponding translations in the target text. This approach may fail in forming source groups if the corpus is too small or if the languages involved are very related and the ambiguity is maintained in the translation, because in the source group there will not be at least another source word to disambiguate from. On the other hand, this procedure can also be applied in absence of a sense inventory in the target language, as long as there is one for the source language (and as long as the sense tags have reference to a inter-linked and rich network such as WordNet).

Diab (2003) tested on an English-Spanish corpus and achieved a recall of 57% on the Senseval-2 English all-words task. Bhattacharya et al. (2004) took inspiration from Diab and Resnik (2002) and Diab (2003) and reformulated their approach in probabilistic terms, reaching a recall of 61% on the same test set. They also proposed a *Concept model* that starts off from the previous model, but tries to overcome language specific senses introducing a concept latent variable that can subsume them. This latter model achieves 65% and also produces a sense inventory for the parallel language as a byproduct.

In following work, Diab (2004) uses the sense annotated data obtained in an unsupervised fashion with SALAAM to train a supervised learning system. The goal was to prove that the range of data such approaches can benefit from is not limited to manually sense-annotated data. The system was tested on 29 nouns in the English lexical sample task of Senseval-2, showing a decisive improvement of 11% over other bootstrapping methods for WSD.

2.1.1 First steps towards CLWSD

Work by Ide et al. (2002) and Tufiș and Ion (2003) on word sense clustering based on translation equivalents shows how crucial the diversity and the number of the languages involved are when tackling the WSD problem in a multilingual perspective; their experiments reached 74% accuracy using six source languages belonging to three different language families, but this figure drops drastically as soon as the variety in languages decreases, since the chances that a polysemous word is lexicalized differently in a different language drops accordingly. Nonetheless, the procedure suggested to find sense distinctions is valuable, as a bilingual lexicon is not required in principle, but can be included if available, with a consequent boost in accuracy.

The knowledge acquired via translation equivalents is in itself useful for other tasks. For instance, Tufiş and Ide (2004) developed a WSD system in order to validate five wordnets for languages in the Balkans area (Bulgarian, Greek, Romanian, Serbian and Turkish), all aligned to Princeton WordNet and developed following the guidelines established for the EuroWordNet project (Vossen et al. 1999)[2]. The tool used translation equivalents to validate the wordnets alignment, looking at the English WordNet first and, if no information could be retrieved from there, backing-off to word sense clustering.

The WSD tool extracts translation equivalents from parallel corpora and uses them to validate the interlingual alignments in the wordnets. For each lexical alignment, the synsets that contain the words in the alignment are retrieved with their InterLingual Index (ILI) codes. At this step, intersection is performed over the two ILI sets to assign the synset that is common to both words, or to find the most semantically related ILI codes otherwise. In this step the correctness of the interlingual alignment is tested and wrong alignments or missing synsets, if any, are detected in the wordnets considered.

The strong point of this method is that even when one wordnet has insufficient information, others can contribute, as further described in Tufiş et al. (2004) and Tufiş and Ion (2004). This procedure can also deal with more complex scenarios: for instance, if there is a tie between two or more ILI pairs, that also happen to have the same relatedness score, the most frequent sense of the target word is selected. This heuristic looks at the English side of the bitext considered, following the idea that word senses obey to a Zipfian distribution and to the *one sense per discourse* heuristic (Gale et al. 1992b).

The back-off mechanism mentioned above (Ide et al. 2002) is applied when

[2] The practice of using Princeton WordNet as an interlingual conceptual representation is a crucial factor in the development of multilingually aligned wordnets.

the current language pair brings no information, but another can, bearing in mind that the more comparable the quality of the wordnets is, the better the back-off strategy will work. Since back-off is performed after attempting to use wordnet information, it is likely that the words to be disambiguated will be clustered together with words that have already been disambiguated. Given this knowledge, the ambiguous words can be labeled with the majority sense in their cluster.

In case no heuristics helps, for instance in the case of a single occurrence of a word that cannot be disambiguated, the sense numbering in Princeton WordNet is weighed in: senses with lower numbers are given preference because, in general, they stand for higher frequencies in a balanced sense-annotated corpus, SemCor (SC) (see Sections 2.2.1 and 3.1.1). 75% accuracy was achieved in disambiguating a portion of *1984* for which there was gold standard annotation. These results, with respect to the granularity level of WordNet 2.0 hierarchy, beat the highest scores back then in monolingual WSD, as the inherent knowledge carried out in the translation process provides decisive clues on the sense intended. In conclusion, the basic methodology is a pioneer example of how to exploit a simple idea such as intersection to obtain clues to sense meaning (see Section 2.2.2).

2.2 Sense annotation using cross-lingual information

Manual semantic annotation is very time-consuming and represents a bottleneck for data-driven NLP systems; as Mana and Corazzari (2002) report, tagging with sense the 80,000 tokens of the SI-TAL Italian Treebank would cost one person's year of work.

In light of the above, there are plenty of parallel corpora at our disposal that could be (at least partially) sense-annotated in an inexpensive fashion by resorting to CLWSD approaches. The translation relation in a multilingual parallel corpus creates a link between words in a translation pair and allows to identify the intended meaning by comparison of the 'semantic baggage' carried by each word.

This section discusses two straightforward, yet powerful, procedures for the sense annotation of parallel corpora, successfully investigated in literature and now taken as a starting point for the present work. All the corpora and sense inventories mentioned below are employed in this contribution, so they are described in further detail in Chapter 3.

2.2.1 Sense projection

The use of sense projection (SP) for automatic annotation was pioneered by Bentivogli and Pianta (2005). The idea of projecting sense annotation is actually present in other nearly contemporary works (e.g. the earlier mentioned Diab (2004)). However, Bentivogli and Pianta investigated more closely the feasibility of porting (manual) sense annotations from one language to another, from both a quantitative and a qualitative standpoint.

Their goal was to create MultiSemCor (MSC), an Italian sense-annotated corpus that was the translational equivalent of the English sense-annotated corpus

SC (Landes et al. 1998), so to provide the community with the first parallel corpus with sense tagging. SC is a subset of the Brown Corpus (Kucera and Francis 1982) of 700,000 running words, of which 200,000 are content words that are enriched with lemma and sense annotation referring to WordNet.

The	discontinuity	can	either	be	that	of	war
	n_10344737			v_01775973			n_10071856
↓	↓	↓	↘	↙	↓	↓	↓
La	discontinuità	può	essere	sia	quella	della	guerra
to	destruction	or	that	of	diplomatic	policy	.
	n_0141128				a_02557914	n_04536028	
↓	↓	↓	↓	↓	↘	↙	↓
di	distruzione	sia	quella	della	politica	diplomatica	.

Table 2.1: Example of sense projection through word alignment on a sentence pair from MSC.

The idea behind this approach is that meaning is generally preserved in the translation process, so the annotations of the English content words can fit the Italian equivalents as well, as long the sense inventory is shared. Table 2.1 exemplifies how the annotation is projected through the word alignment.

Bentivogli and Pianta carried out a pilot study starting from six SC texts in order to investigate the feasibility of this approach (Bentivogli and Pianta 2002) and, especially, to determine whether the English-Italian word aligner KNOWA, developed with this very study in mind, would be sufficiently good or not. The preliminary results being promising, Bentivogli and Pianta went on applying the annotation transfer methodology on a larger scale.

After selecting 116 texts from SC, the first step was having their Italian translations produced by professional translators. Upon this first manual step, the parallel corpus was created by sentence- and word-aligning the two sides. At this point, the actual annotation transfer is very easy: the English annotations were copied to their aligned Italian translations by exploiting word alignment as a bridge.

As for the shared sense inventory, the choice fell on MultiWordNet (MWN)[3] (Pianta et al. 2002), a multilingual WordNet with reference to WN 1.6. One of the research goals on the table was to ascertain if such a parallel corpus could be used to validate and test the coverage of this multilingual sense inventory, and eventually automatically enrich it with any missing word senses detected during the process.

To better assess the feasibility of the transfer procedure in all its steps, a gold standard of four texts was formed. Each text was translated twice: *free* translations were produced with no specific recommendations, while *controlled* translations were following specific guidelines aimed to maximize the word alignment. This was done to assess the practicality of this approach - and, in particular, the alignment precision - to already existing parallel corpora, built with no such application in mind. All eight translations were hence aligned manually, and annotators were asked to align different types of units.

The four controlled translations were also manually annotated with sense: the

[3] http://multiwordnet.fbk.eu/

annotators could look at the original SC texts for reference, but they were asked to select a different sense in case the English annotation was not considered appropriate. Throughout the process, they had to mark different kinds of semantic correspondence between the aligned units.

Table 2.2 reports the results of the evaluation on the gold standard. Given 4,313 Italian words to be annotated in the gold standard texts (from 4,101 original English annotations), Bentivogli and Pianta achieved a precision of 87.9% and coverage of 76.4%, thus leaving 24.6% of the content words to be annotated. These results were considered satisfactory, as the human effort in correcting the errors and completing the annotation would be, in any case, greatly reduced compared to annotating the corpus from scratch. In addition, the method provides a new parallel corpus as a by-product, which is already beneficial in itself.

	Precision	Recall	Coverage
Italian controlled texts	87.9	67.2	76.4

Table 2.2: Evaluation of the Italian annotation. Readapted from Bentivogli and Pianta (2005)

As shown, 12.1% of the transferred annotations transferred were wrong. Breaking this figure down to the possible sources of error, non-transferable annotations caused almost half of it (5.9%); 3.3% is due to wrong source annotations: almost all the wrong source annotations marked by the annotators in the gold standard were transferred (109/117). The last 2.9% depends on wrong alignments. This last figure proves that the KNOWA was very precise, although the peculiarities of the task facilitated the alignment step, as further discussed in Section 3.1.2.

The gold standard text g-11 was manually annotated in both its controlled and free translations in order to assess the annotation transfer in a real case scenario. While the scores of the controlled translation align to the general evaluation, the free translation scores clearly show that the task got more difficult, being 84.8% precision (2.9% drop compared to the controlled translation), 63.1% recall (drop of 7.7%) and 74.4% coverage (drop of 6.3%).

Ultimately, it cannot be stressed enough how much the validity of SP relies on the precision score of the word alignment. Table 2.3 summarizes the evaluation of KNOWA for both free and controlled translations, with respect to English words with sense annotation.[4] Controlled translations allow, as expected, better alignment, which even improves when only content words are considered.

The most important figure is perhaps the 94.7 precision score in controlled translations, only looking at the sense tagged words. The error rate of 5.3% is actually lower if only alignments that lead to wrong annotation transfer are considered; this explains why the alignment error part of the 12.1% error rate is lower. Error analysis revealed that adverbs and adjectives represent a tougher challenge for KNOWA, thus making the recognition of multiwords containing these lexical categories an actual goal for improving the procedure.

To better understand MSC goals, a comparison with related work by Diab and

[4] In the release documentation, precision and recall values for both full-text and sense-tagged-only controlled evaluation are actually 0.1% higher.

	Translation	Precision	Recall	Coverage
Full-text	Free	85.9	61.8	70.0
	Controlled	**89.2**	**69.4**	**76.1**
Sense tagged words	Free	92.8	70.3	75.8
	Controlled	**94.7**	**76.2**	**80.5**

Table 2.3: Performance of KNOWA on full-text and on sense-tagged words only, for both free and controlled translation. Readapted from Bentivogli and Pianta (2005)

Resnik (2002) might be noteworthy. This unsupervised method allows the sense annotation of nouns in a parallel corpus formed by the Brown corpus and its automatic translations in German, French and Spanish. Translations are exploited as hints to WSD but, unlike SP, this procedure does not require a text in the corpus to be already annotated with sense, so it is actually easier to apply. At the same time, Diab's method has different goals in terms of the quality and the target of annotation; to start with, it limits itself to the annotation of nouns, and so it can perhaps handle better the fact that translation is automatically produced by a MT system and alignment is taken care of by a language-independent statistical aligner. On the other hand, SP by Bentivogli and Pianta aims to annotate all content words and decisively counts on very precise word alignment and source text annotation, as its ultimate goal is to provide new sense-annotated texts for training of WSD systems.

SP represents a valid strategy to relieve the knowledge acquisition bottleneck, as it creates motivation to produce translations of existing corpora and, consequently, new parallel corpora (a valuable resource in itself). Moreover, SP can potentially exploit the availability of sense-annotated resources in highly-represented languages such as English to bootstrap the creation of similar resources in less researched languages, eliminating the need for manually annotating the translation itself. The satisfying precision and coverage scores of MSC encouraged others to pursue similar work (Bond et al. 2012; Lupu et al. 2005, see Chapter 3).

The main argument against the reproducibility of the experiment is that sense-annotated corpora, if available at all, are not usually translated in other languages, so there is certainly lack of corpora to which apply the cross-lingual transfer. The NTU-MC corpus (Tan and Bond 2014) is one of the few exceptions and could make a good candidate.

Europarl (Koehn 2005), while not manually sense-annotated, is another good resource for work on CLWSD, as it is a professionally translated parallel corpus consisting of the aligned proceedings of the European Parliament in 21 European languages.

On the other hand, while there are many methods to automatically annotate a corpus, it would make little sense to project sense annotations that have not been checked by human annotators, as supervised approaches to WSD require high-quality data. As a result, sense projection is a valid, but often infeasible option.

2.2.2 Sense intersection

As observed, sense projection requires at least one side of the parallel corpus to be manually annotated in order to export annotations to the available translations. Gliozzo et al. (2005) had the merit of first coming up with the idea of using aligned wordnets and sense intersection as a means to automatically sense annotate at once both texts in a parallel corpus. This procedure, although unsupervised, relies on the reasonable expectation that one can identify the sense actually intended in context by looking at the *polysemic differential* between two languages. This gives very good promise of bootstrapping sense annotations that are actually correct and safe to use for training supervised WSD systems. Their work differs from Tufiş and Ide (2004) both for the type of application and for the extensive quantitative and qualitative testing on English and Italian.

This approach exploits the fact that MWN is a multilingually aligned lexical resource: for any aligned translation pair, the set of senses of each word can be retrieved and one can disambiguate the translation pair by simply intersecting the two sets of senses retrieved. If the intersection only leads to a single sense in common, then the words are fully disambiguated, which is the ideal case. Otherwise, the ambiguity remains, but it is greatly reduced in most cases, making the intervention of a human annotator less time-consuming.

Given the availability of wordnets that have high coverage and are well aligned (i.e., the requirement that all sense distinctions present in one language are also in the other is satisfied), the method manages to disambiguate 51% of MultiSemCor

(Bentivogli and Pianta 2005) with 100% precision. In real scenarios, however, this figure drops to 67% precision and 41% coverage, because of alignment errors, interlingual differences and, above all, insufficient coverage in the Italian WordNet (Pianta et al. 2002), far less developed than Princeton WordNet.

Despite the limits due to the different development stages, generally speaking the availability of an aligned multilingual lexical resource allows to inexpensively annotate parallel text by simply assigning all words their sets of senses and then performing intersection to find the subset that is shared in different languages, in any case reducing the ambiguity. In fact, polysemy reduction in the all-words scenario is striking in both languages: the average number of senses per word decreases to 1.54, starting from an average number of senses of 6.72 in English and 3.28 in Italian. In the authors' words, this method simply exploits the "polysemic differential between two languages".

The procedure hereafter named **MSI**, whose discussion is postponed to Chapter 4 and which represents the core of this work, is substantially very similar to Gliozzo et al. (2005), but it is set in a multilingual scenario and, consequently, uses different heuristics to attempt all-words disambiguation. The state of the resources employed is very different as well, as more lengthily discussed in 3.2.

2.2.3 Recent developments in CLWSD research

Especially after the introduction of a dedicated task in SemEval-2013 (Lefever and Hoste 2013), work on CL-WSD has increased, driven by the increasing evidence that multilingual translational knowledge is more informative than only relying on knowledge induced by monolingual text.

Lefever and colleagues, among others, have been particularly prolific in this field in recent years. In Lefever et al. (2011), and more in detail in Lefever (2012), they presented ParaSense, a machine learning and multilingual approach to WSD that does not rely on resources such as WordNet, but tries instead to derive word senses from parallel corpora, as already successfully attempted in the literature (Diab 2003; Ide et al. 2002; Ng et al. 2003). The novelty is that their approach is classification-based, building one classifier for each target language: the aligned translations contribute directly to disambiguate the target language via information computationally represented in feature vectors, which include both English local context features and binary translation features extracted from the aligned translations.

The system outperformed the classifiers that only use local context information in the 'Cross-Lingual Word Sense Disambiguation task' in SemEval-2010 for the French, Dutch, Spanish and German languages (Lefever and Hoste 2013). In Lefever et al. (2013), the system outperformed the other competitors for all five languages, thus also including Italian. What makes this method applicable to languages that do not have high-quality, good-coverage lexical resources is that all the information used is extracted from the parallel text. In addition, the framework is completely language-independent.

2.2.4 Breathing new life in cross-lingual sense annotation

As for the present work, the encouraging results from Bentivogli and Pianta (2005) and Gliozzo et al. (2005), while certainly not the sole answer to the CLWSD problem, suggest valid intuition about meaning and prove to be useful for a community that is still struggling with the lack of sense-annotated corpora. Adding translational evidence from multiple languages holds great promise for higher coverage and precision than using only monolingual or bilingual information, as proved by preliminary studies on MSI (Bonansinga and Bond 2016; Bond and Bonansinga 2015).

After an overview of the resources and the requirements needed in the following chapter, MSI is exhaustively described and evaluated in Chapter 4.

Chapter 3

Multilingual sense intersection: resources and requirements

This chapter describes in detail the resources used in this thesis. It also lists the preprocessing steps that need to be followed to carry out the multilingual sense intersection (MSI) procedure described in section 2.2.2.

3.1 Corpora

3.1.1 SemCor

SemCor (SC) (Landes et al. 1998) is a sense-annotated corpus developed at Princeton University from a selection of texts from the Brown Corpus of Standard American English (Kucera and Francis 1982). The Brown corpus was the first computer readable corpus of general content of American English and it consists of 500 texts, each around 2,000 words long, to a total of one million words.

The texts in the Brown corpus were chosen so that the collection is balanced with respect to 15 different genres and literary styles: press (and sub-genres: reportage, editorial and reviews); religion, skill and hobbies, popular lore, belles-lettres, miscellaneous: government and house organs, learned, fiction (general; mistery; science, adventure; romance) and humor.

As Landes et al. (1998) report, around 80% of the content words in the Brown Corpus are polysemous, because frequently used words tend also to have multiple meanings. Consequently, the task of semantically tagging part of it was very laborious, as also showed by the low value of inter-annotator agreement of 78.6%,

	Texts	Tokens	Annotated tokens
All-words	186	359,732	192,639
Only verbs	166	316,814	41,497

Table 3.1: Statistics of SemCor all-words and only-verbs components.

The Princeton corpus, named SemCor after *sem*antic con*cor*dance, consists of 352 texts, of which 186 have lemma, PoS and semantic annotation for all content words (nouns, verbs, adverbs and adjectives), while in the remaining 166 texts only verbs are annotated with lemma and sense; see Table 3.1. Compared with current corpus standard, SC is nowadays considered very small. However, the fact that it is one of the few semantically annotated resources still makes it valuable.

The first preprocessing step was to convert the chosen extracts in a standard SGML text file, so to properly encode paragraph, sentence and word segmentation. Then, tokenization and PoS tagging with the Brill tagger (Brill 1992) were performed.

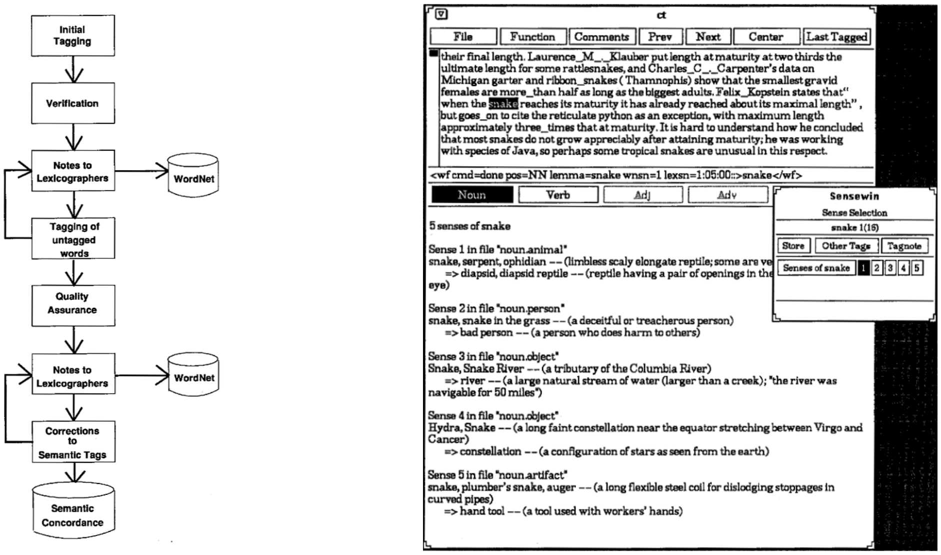
```
<contextfile concordance=brown>
<context filename=br-a01 paras=yes>
<p pnum=1>
<s snum=1>
<wf cmd=ignore pos=DT>The</wf>
<wf cmd=done rdf=group pos=NNP lemma=group wnsn=1 lexsn=1:03:00:: pn=
   group>Fulton_County_Grand_Jury</wf>
<wf cmd=done pos=VB lemma=say wnsn=1 lexsn=2:32:00::>said</wf>
<wf cmd=done pos=NN lemma=friday wnsn=1 lexsn=1:28:00::>Friday</wf>
<wf cmd=ignore pos=DT>an</wf>
<wf cmd=done pos=NN lemma=investigation wnsn=1 lexsn=1:09:00::>
   investigation</wf>
<wf cmd=ignore pos=IN>of</wf>
<wf cmd=done pos=NN lemma=atlanta wnsn=1 lexsn=1:15:00::>Atlanta</wf>
<wf cmd=ignore pos=POS>'s</wf>
<wf cmd=done pos=JJ lemma=recent wnsn=2 lexsn=5:00:00:past:00>recent</
   wf>
<wf cmd=done pos=NN lemma=primary_election wnsn=1 lexsn=1:04:00::>
   primary_election</wf>
<wf cmd=done pos=VB lemma=produce wnsn=4 lexsn=2:39:01::>produced</wf>
<punc>``</punc>
<wf cmd=ignore pos=DT>no</wf>
<wf cmd=done pos=NN lemma=evidence wnsn=1 lexsn=1:09:00::>evidence</wf>
<punc>''</punc>
<wf cmd=ignore pos=IN>that</wf>
<wf cmd=ignore pos=DT>any</wf>
<wf cmd=done pos=NN lemma=irregularity wnsn=1 lexsn=1:04:00::>
   irregularities</wf>
<wf cmd=done pos=VB lemma=take_place wnsn=1 lexsn=2:30:00::>took_place
   </wf>
<punc>.</punc>
</s>
</p>
```

Listing 3.1: An extract from SC text br-a01.

As word forms in a sentence are all tagged by default with wf, this format is convenient because the collection may also include text that is untagged or only partially tagged. In the extract in Listing 3.1 it is showed how a sentence is tagged in SC. The original release was annotated with reference to WN 1.6, and the semantic tagging served also the purpose of testing the coverage of WN; as the annotators proceeded, any missing senses and words were included in the lexical base.

Together, the `lemma` and the `lexsn` values permit an unique sense reference to the WN 1.6 database. This *sense key encoding* embodies a variety of information: the synset type (noun, verb, adverb, adjective or satellite adjectives), the lexicographer file number containing the synset and the `lex_id` that uniquely identifies a sense. When the sense belongs to an adjective satellite synset, the last two positions of the sense key are filled with the head adjective for the satellite synset and its `lex_id` within a lexicographer file. Sense keys are not the only method to point to WN senses, but are recommended by Princeton WordNet (PWN) developers because they are stable across different versions of the database.

Fig. 3.1a shows the semantic tagging process followed by SC creators. The automatic procedure presents one word at a time to the tagger and allows them to select a WN sense for every open-class word, as showed in Fig. 3.1b. If a sense is missing from WN or is found duplicate of another, the interface allows the tagger to add a note, which will be automatically sent to the lexicographers in charge. The process is iterative, in that taggers and lexicographers cooperate until they agree on sense distinctions to be made.



their final length. Laurence_M_._Klauber put length at maturity at two thirds the ultimate length for some rattlesnakes, and Charles_C_._Carpenter's data on Michigan garter and ribbon_snakes (Thamnophis) show that the smallest gravid females are more_than half as long as the biggest adults. Felix_Kopstein states that" when the snake reaches its maturity it has already reached about its maximal length" , but goes_on to cite the reticulate python as an exception, with maximum length approximately three_times that at maturity. It is hard to understand how he concluded that most snakes do not grow appreciably after attaining maturity; he was working with species of Java, so perhaps some tropical snakes are unusual in this respect.

<wf cmd=done pos=NN lemma=snake wnsn=1 lexsn=1:05:00::>snake</wf>

5 senses of snake

Sense 1 in file "noun.animal"
snake, serpent, ophidian -- (limbless scaly elongate reptile; some are ve
=> diapsid, diapsid reptile -- (reptile having a pair of openings in the
eye)

Sense 2 in file "noun.person"
snake, snake in the grass -- (a deceitful or treacherous person)
=> bad person -- (a person who does harm to others)

Sense 3 in file "noun.object"
Snake, Snake River -- (a tributary of the Columbia River)
=> river -- (a large natural stream of water (larger than a creek); "the river was navigable for 50 miles")

Sense 4 in file "noun.object"
Hydra, Snake -- (a long faint constellation near the equator stretching between Virgo and Cancer)
=> constellation -- (a configuration of stars as seen from the earth)

Sense 5 in file "noun.artifact"
snake, plumber's snake, auger -- (a long flexible steel coil for dislodging stoppages in curved pipes)
=> hand tool -- (a tool used with workers' hands)

(a)

(b)

Figure 3.1: The workflow followed to build SemCor (a) and the annotation interface (b). From Landes et al. (1998, pp. 204, 206)

3.1.2 MultiSemCor

MultiSemCor, as seen in Section 2.2.1, was built through sense projection. The main advantage of this approach is to use existing resources to bootstrap the creation of new ones: MSC is aligned at the word level and annotated with part of speech, lemma and word sense, and the Italian corpus can be used independently.

The procedure and the quality of the resulting annotation have already been discussed; hence, the present section gives deeper perspective on the requirements and the issues that arose in the making of the parallel corpus.

The pilot study - As already mentioned, the actual making of MSC was preceded by a pilot study on six texts (some of which were later used to form a gold standard), to a total of 12,000 English running words. The texts were chosen so to represent both the Imaginative Prose (Brown corpus' texts `l-10` and `p-12`) and the Informative Prose (texts `b-13`, `f-03`, `g-11`, `j-53`) sections from the Brown corpus (Kucera and Francis 1982).

Four texts were translated twice as a free and a controlled translation, for which the annotators were asked to consult preferably the same dictionaries used to feed the word aligner, so to maximize the lexical correspondences. The guidelines recommended that translators maintained the same sentence segmentation, marked Italian multiwords and named entities following SC's conventions and chose the nearest translation equivalents, trying to maintain the same PoS as well. However, the guidelines made it clear that a fluent, natural-sounding Italian translation held priority over the controlled translation criteria.

Free translations were to check how well the alignment (and, consequently, the annotation transfer) would deal with a regular free translation, compared with one designed to facilitate the task. In the assignment step, the word aligner also selects lemmatization and PoS for the word to be aligned: compared to the gold standard, the automatic alignment scored 91% precision. Text g-11 was aligned by two annotators in both its free and controlled translations, respectively totaling an inter-annotator agreement of 87% and 92%.[1]

Finally, as seen in Chapter 2, the gold standard texts were manually aligned in order to be able to assess the transfer procedure. Semantic annotation is far from being trivial and requires specific training, so to give assurance that the task will be addressed consistently and coherently with the given guidelines. In addiction, semantic judgments are, generally speaking, exposed to subjectivity and sense distinctions in the reference sense inventory may be so fine-grained that different annotators might give different tags, both being suitable for the target word. The preliminary training that annotators were given helps restrain such subjectivity, but adds further time and money requirements. Overall, the inter-annotator agreement, computed with the Dice coefficient method, amounts to 81.9%, which is higher than the one resulting from the original SemCor annotation (Fellbaum 1998).

The resulting gold standard includes 8,877 English tokens and 9,224 Italian

[1] The agreement rate is the Dice coefficient defined in Véronis and Langlais (2000), computed as the fraction of number of the units selected by both annotators and the number of units in the texts.

tokens in controlled translations, i.e. Italian texts contain 3.9% more tokens. This is due to the characteristics of the language, which is abundant in clitics and articles. The difference in size also concerns the annotated words, because in Italian modal and auxiliary verbs and partitives were annotated, differently from English.

Word alignment - The English-Italian word aligner KNOWA (KNOwledge-intensive Word Aligner) was employed for the word alignment. KNOWA exploits information extracted from the Collins bilingual dictionary and also a morphological analyzer and a multiword recognizer for both languages (Pianta and Bentivogli 2004).

As preprocessing steps, for each English and Italian word a set of candidate lemmas is produced and sorted by probability looking at the part of speech. Then, the actual alignment phase begins, taking in input a sentence pair at the time.

The alignment procedure is incremental and is based on finding potential correspondences between English and Italian words: if two candidate lemmas happen to be translation equivalents of each other, then the respective tokens are aligned. In more detail, the procedure is firstly done from English to Italian and targeting the Italian words whose position is within the window given by the English word position ± a certain value, specified as parameter. This poses a problem, as longer sentences may have more than one potential correspondent, but the algorithm stops searching when it finds the nearest one, which may not always be the correct one.

Once the algorithm has attempted to align every English word, it starts again in the opposite language direction. Finally, if any unaligned words are left, the algorithm tries to pair them resorting to their graphemic similarity. At this stage, the annotation transfer is simply done by assigning the sense annotation from the English to the Italian words selected by the alignment.

Admittedly, the union of KNOWA and SemCor is particularly blessed for at least three reasons; first, words were lemmatized and POS-tagged and then manually checked, preventing errors from multiple potential matches. Secondly, in SemCor multiwords (which usually give word aligner systems a hard time) are already marked as such when present in WordNet. The setting would not be as ideal with other texts, because the only other source of knowledge, the Collins dictionary, only covers a limited part of the multiwords actually used in texts. Finally, the aim of this experiment is to produce annotation for content words, which are easier to align than functional words. If alignment is made easier, then the annotation based on it benefits accordingly. MultiSemCor, as seen in Section 2.2.1, was built through sense projection. The main advantage of this approach is to use existing resources to bootstrap the creation of new ones: MSC is aligned at the word level and annotated with part of speech, lemma and word sense, and the Italian corpus can be used independently.

The procedure and the quality of the resulting annotation have already been discussed; hence, the present section gives deeper perspective on the requirements and the issues that arose in the making of the parallel corpus.

Theoretical issues on annotation - Bentivogli and Pianta discuss in detail a range of theoretical issues to weigh when working with translations. First of all, lexica of different languages are not fully comparable; in the case of English and Italian, Bentivogli and Pianta (2000) found that in MultiWordNet 7.8% of the English words correspond to lexical gaps in Italian.

A second issue is that translated texts only partially represent current language: they tend to be less ambiguous, less figurative and more conventional in their stylistic choices, thus not making very representative exemplars of their language. Being SemCor composed of written, formal text, Bentivogli and Pianta considered this latter not being much of a real argument.

Quality issues on annotation - As for quality issues, the procedure includes many possible sources of error with respect to the annotation transfer. First, while SemCor was tagged manually, there are admittedly cases in which the wrong annotation was assigned. Also the alignment might be incorrect and introduce annotation errors.

There are also cases in which annotations could not be projected because the translation equivalents in another language are not cross-language synonyms, even though the chosen translation works better in context than a literal one, which would make the sentence sound unnatural. Likewise, sometimes another language requires some rewording to the point that the overall sentence meaning is preserved, but at the word level the cross-lingual synonymity has been lost, thus making the alignment - and, consequently, any annotation transfer - incorrect.

Lexical gaps constitute another special case, in which an English word does not have a single-word correspondent in Italian, but instead the meaning can only be realized with a free combination of words. In principle, each component of such combination should be tagged with its own sense. However, for evaluation purposes, only the lack of synonymity at the lexical level was regarded as an annotation error. 16.9% of the English annotations were non-transferable. Of these, 85.4% were due to lack of cross-language synonymity, while the remaining 14.6% were due to translation equivalents that are not lexical units.

It should be noted that KNOWA, being based on a bilingual dictionary, is less likely to align non-synonymous translation compared to statistical word aligners, thus making this tool more suitable for this type of task. In fact, only 33.2% (196/591) of the wrong translations in the gold standard have been aligned by KNOWA.

MultiSemCor release - MultiSemCor 1.0 was released in January 2005 and since then has been available for download upon request on the official website.[2] The corpus consists of 116 English texts from SemCor aligned to their Italian translations, to which English annotations were projected to. Both sides are annotated with paragraph and sentence splitting, morphosyntactic information and word sense information with reference to WordNet 1.6. The release also includes the alignment files, which mark both sentence and word alignment.

The corpus is encoded in XML, with special regard for the Corpus Encoding Standard guidelines.[3].The English files were ported to this format, but contain all

[2] http://multisemcor.fbk.eu/

[3] See http://www.cs.vassar.edu/CES/ and http://http://www.xces.org/

the information present in the original SGML files.

Among the 116 Italian files, seven (`b-13`, `f-03`, `f-43`, `g-11`, `j-53`, `l-10` and `p-12`) were controlled translations that were aligned manually to serve as gold standard (`crit-gs`). `b-13` and `f-03` were also translated with a mixed approach, i.e. the translators did not follow any guidelines, but knew that the text was to be used in an automatic word alignment task. No free translations were released. As for the morphosyntactic annotation, for Italian files it was carried out using tools developed by Bentivogli's and Pianta's research group.

In the English files, each token has the tagging status, which may be `done` (indicating a content word to be annotated), or `ignore` (indicating a functional word). In the Italian files the tagging status may be: `transfer` when the word was successfully aligned and the English sense transferred; `no-transfer` if the word was not aligned, so no sense can be selected from an English counterpart; `ignore` when the translation pair was aligned, but the original English tagging status was `ignore`, so there is again no sense to transfer. In both languages, finally, proper names have their type specified in the lemma value, with the following options: *person*, *place*, *group*, *location* or *other*.

3.1.3 Romanian SemCor

Inspired by the work of Bentivogli and Pianta, Lupu et al. (2005) developed a Romanian SemCor with the purpose of enriching MultiSemCor by providing Romanian translations for all the 116 English texts already included, planning of aligning the Romanian and Italian sides near in the future. This extended project bears the name of **MultiSemCor+**.

While the Romanian SemCor itself has been further developed even since, the status of the Romanian component in the MultiSemCor website has remained the same, i.e. it still includes only 12 texts in Romanian, aligned to their English counterparts. Lupu et al. (2005) report that 22 more texts were in progress of alignment already back then, making a total of 65,926 tokens in 3,871 sentences.

Tufiş et al. (2004) used the tool RACAI to segment and sentence-align the corpus. RACAI can also identify and properly handle complex nouns, phrasal verbs, idioms and other multiword expressions. A row tagging method with combined language models (Brants 2000; Tufiş 1999) came to the aid of tagging and lemmatization. The sense annotations were the same as in SemCor, with reference to WN 2.0. For this reason, it was necessary to map the annotations to the WN 1.6 sense inventory that is instead adopted in MultiSemCor.

Summing up, the Romanian component of the MultiSemCor consists of the translations of 34 English SemCor texts, each tagged with paragraph and sentence splitting information and morphosyntactic and sense annotation. Of this set, the following 12 texts are aligned and actually available online: `br-e23`, `br-e27`, `br-e28`, `br-e30`, `br-f14`, `br-f15`, `br-f16`, `br-f22`, `br-g43`, `br-h17`, `br-j29`, `br-j34`.[4]

As it is distributed, Romanian SemCor (RSC) is unfortunately not word-aligned to any other component of the parallel corpus, which is a requirement to perform

[4] http://multisemcor.fbk.eu/frameset1.php

sense mapping with any of the mentioned procedures in Section 2.2. Nevertheless, as the sentence alignment is available and as we are only interested in content words, a tentative word alignment was attempted based upon the information already available.

The English-Romanian parallel corpus

As mentioned above, the English-Romanian parallel corpus has been further developed, independently of the MultiSemCor project. Ion (2007) reports that it consists of 178,499 words for English and 175,603 words for Romanian. The corpus is made up of 81 manually translated texts,[5], of which 50 are shared with MSC, and annotated followingly WN 3.0. See Appendix 6 for the list of the 50 texts shared by all SemCor corpora.

Ion (2007) and Tufiş et al. (2008) describe the tool used to automatically assign lexical and morphosyntactic information, the TTL tagger, which is reported of having achieved 98% of accuracy. The annotation provided by this tool is very descriptive and detailed. Each token is tagged with its base form and a complex MSD tag, a rich description that includes the grammatical meta category, chunk information and sense annotation referring to WN 3.0 sense inventory. The MSD convention follows the specifications described in Erjavec (2004)[6]: as many as 614

[5] Actually 82 in the release.

[6] The tag conventions used are available online at http://nl.ijs.si/ME/V4/msd/html/msd.msds-ro.html (for Romanian) and at http://nl.ijs.si/ME/V4/msd/html/msd.msds-en.html (for English).

MSDs are available for Romanian.

The parallel corpus consists of two XML files heavily annotated with information on paragraph, sentence, constituent group and word structure is given, following the XCES format 3 described by Ide et al. (2000). The electronic resource comes with detailed documentation about the tools used for the preprocessing and the annotation; in addition, it is distributed with a 'Non Commercial No Redistribution No Derivatives' license (NC-NoReD-ND) and is freely available on THE META-SHARE [7] platform for research purposes upon filling of the license.

3.1.4 Japanese SemCor

Bond et al. (2012) built Japanese SemCor (JSC) with the goal of providing Japanese counterparts for the texts covered in MSC. As a preprocessing step, English WN 1.6 sense annotations were ported to WN 3.0 using the mappings provided by Daude et al. (2003). The alignment was carried out manually, but some automatic post-processing was performed before attempting the annotation transfer; in particular, the peculiarity of English-Japanese translation led to allow many-to-many word alignments.

Early on in the translation process, translators expressed their concern about a potential lack of consistency given by the fact that the sentences to be translated had been assigned out of sequence. Upon this complaint, contiguous sentence blocks were assigned instead, with the option of both looking at the last ten trans-

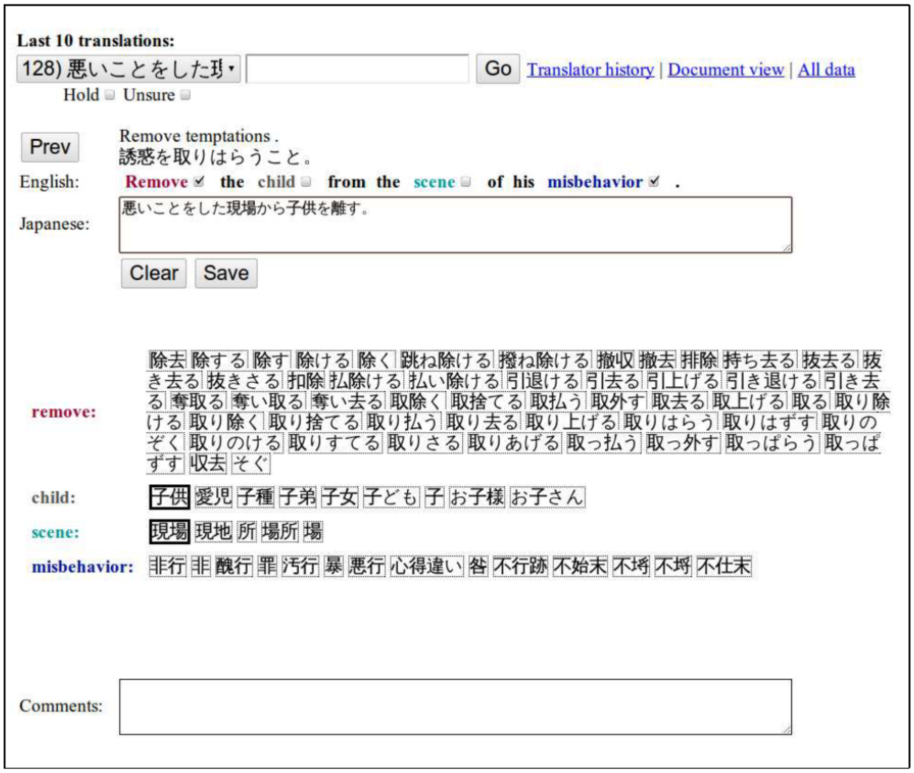
[7] http://metashare.elda.org/

lated sentences and at the document overall to repair for the initial translation conditions.

In particular, the guidelines recommended that translators: conform to formal Japanese (except for clearly informal texts, of course); try to determine the canonical translation of proper names; maintain the same sentence segmentation as in the original English, except for translations resulting in clearly stilted Japanese; reorder the words and include discourse connectors when needed, to the benefit of better readability.

Sense annotation was carried out through sense projection by exploiting the word alignment, similarly to what was done for Italian. This work differs from MSC in that Japanese is less close to English than Italian is; for this reason, translators were provided with sense-specific translations to potentially use to boost coverage. Fig. 3.2 is a screen shot of the annotation interface: translators could consult the sense-specific annotations suggested, comment, mark unsure translations or leave them for later.

As in MSC, some WN 1.6 have been dropped in WN 3.0, so the affected annotations are not available anymore. In the end, 39% of senses could be transferred; this figure would increase of 9% if the missing terms and senses detected throughout the translation (13,857 cases) were added to Japanese WordNet (JWN). Most cases of lexical gaps are due to the frequent part-of-speech mismatches occurring with translation: for example, many Japanese words are just not listed in the JWN because they are predictably compositional, so it would be conceptually wrong to



128) 悪いことをした現

Hold □ Unsure □

Remove temptations .
誘惑を取りはらうこと。

English: Remove ☑ the child □ from the scene □ of his misbehavior ☑ .

Japanese: 悪いことをした現場から子供を離す。

remove: 除去 除する 除す 除ける 除く 跳ね除ける 撥ね除ける 撤収 撤去 排除 持ち去る 抜去る 抜き去る 抜きさる 扣除 払除ける 払い除ける 引退ける 引去る 引上げる 引き退ける 引き去る 奪取る 奪い取る 奪い去る 取除く 取捨てる 取払う 取外す 取去る 取上げる 取る 取り除ける 取り除く 取り捨てる 取り払う 取り去る 取り上げる 取りはらう 取りはずす 取りのぞく 取りのける 取りすてる 取りさる 取りあげる 取っ払う 取っ外す 取っぱらう 取っぱずす 収去 そぐ

child: 子供 愛児 子種 子弟 子女 子ども 子 お子様 お子さん

scene: 現場 現地 所 場所 場

misbehavior: 非行 非 醜行 罪 汚行 暴 悪行 心得違い 咎 不行跡 不始末 不埒 不埒 不仕末

Comments:

Figure 3.2: Example of the annotation interface made available to the translators. From Bond et al. (2012, p. 58)

include them in a dictionary. The authors considered to solve this issue by including pertainym links to the PWN structure in future work.

JSC was encoded with respect to the Kyoto Annotation Format (KAF) (Bosma et al. 2009) and is released under the same license as SC.[8] The corpus consists of 14,169 sentences with 150,555 content words, of which 58,265 are sense tagged with annotations automatically transferred from English (131 of them with more than one sense).

Listing 3.2 shows a sample KAF record. The reference to SC file and sentence id is recorded and each token `wf` is further annotated with a lemma and pos and WN sense annotation if available through the element `term`.

This project is closely intertwined with JWN, in that it both provides sense frequency data and is used to detect missing senses, so to eventually enrich the lexical base. Moreover, the nature of the project allows to think in terms of a trilingual sensebank, that can be exploited for a variety of applications in translation.

```
<?xml version="1.0" encoding="utf8"?> <KAF lang="jpn"> <kafHeader> <
    fileDesc filename="br-k01"/> <linguisticProcessors layer="text"> <lp
     timestamp="2011-09-23T11:45:18" version="0.98" name="MeCab"/> </
    linguisticProcessors>
</kafHeader> <text> <wf wid="w1.1.1" sent="1" para="1">スコッティ</wf>
    <wf wid="w1.1.2" sent="1" para="1">は</wf> <wf wid="w1.1.3" sent="1"
     para="1">学校</wf> <wf wid="w1.1.4" sent="1" para="1">に</wf> <wf
    wid="w1.1.5" sent="1" para="1">戻ら</wf> <wf wid="w1.1.6" sent="1"
    para="1">なかっ</wf> <wf wid="w1.1.7" sent="1" para="1">た</wf> <wf
    wid="w1.1.8" sent="1" para="1">。</wf>
</text> <terms> <term tid="t1.1.1" lemma="スコッティ" type="open" pos="
    N.名詞.一般"> <span> <target id="w1.1.1"/>
</span> <component lemma="スコッティ" id="c1.1.1" pos="N.名詞.一般"/>
```

[8] Both the Japanese WordNet and the Japanese SemCor are available at the following address: http://compling.hss.ntu.edu.sg/wnja/index.en.html

```
</term> <term tid="t1.1.3" lemma="学校" type="open" pos="N.n"> <span> <
    target id="w1.1.3"/>
</span>
<target id="w1.1.3"/>
</span> <component lemma="学校" id="c1.1.3" pos="N.名詞.一般"/> <
    externalReferences> <externalRef resource="Wordnet jpn 1.1"
    reference="jpn-11-学校-n"/> </externalReferences>
</term> <term tid="t1.1.5" lemma="戻る" type="open" pos="V.v"> <span> <
    target id="w1.1.5"/>
</span> <component lemma="戻る" id="c1.1.5" pos="V.動詞.自立"/> <
    externalReferences> <externalRef resource="Wordnet jpn 1.1"
    reference="jpn-11-戻る-v"/> </externalReferences>
</term> </terms>
</KAF>
```

Listing 3.2: Sample KAF record for スコッティは学校に戻らなかった。(Scotty ha gakkō ni modora nakat ta .), the Japanese translation of English sentence *Scotty did not go back to school.*. Readapted from Bond et al. (2012, p. 62).

3.1.5 The multilingual corpus used in the present work

From here on, the multilingual corpus built from the shared parts of the SemCor-based corpora will be referred to as Multilingual Parallel Corpus (MPC). MSI is applied on this corpus, as described in the next chapter.

Unfortunately, only 49 texts are available in English, Italian, Romanian and Japanese at the same time, as the RSC project did not translate the 116 texts originally chosen in MSC, but a different subset of the original SC, of which only 50 texts were shared with the other projects. The common set is further reduced by one due to the fact that the Japanese release hereby used misses text d02.

Table 3.2 sums up the basic statistics of each SemCor corpus. In the case of English and Italian, the number of target words after the migration to WordNet 3.0

	Texts	Tokens	Content words	After mapping
EM	116	258,499	119,802	118,750
IT	116	268,905	92,420	92,022
RO	82	175,603	48,634	=
JP	116	382,762	150,555	58,265

Table 3.2: Statistics for each component of the multilingual parallel corpus built from SemCor.

(WN 3.0) is also specified.

No language in the MPC comes already word-aligned with all the others. English SC is aligned to Italian and Japanese texts, but it only has sentence alignment with the Romanian corpus. The other three languages have no connection, having been developed primarily to double the English SC. Consequently, word alignments for each pair need to be produced, as they are a necessary requirement to perform sense mapping.

Despite the MPC consisting of 49 texts only, the original SC corpora can offer, depending on the language pair, several more text pairs that can come to increase the training data. Table 3.3 gives a clearer picture of the aligned sentences and texts available for each language pair.[9] Section 3.3.2 discusses in detail the steps taken to provide the word alignment for each pair.

[9] In RSC release, text `j-22` only consists of 5 sentences, so it would be more accurate to say that EN-RO parallel corpus from MSC consists of 81 texts.

Language pair	Sentences	Texts
EN-IT	12,842	116
EN-RO	8,276	82
EN-JP	12,781	115
IT-RO	4,974	50
IT-JP	12,781	115
RO-JP	4,913	49

Table 3.3: Sentences available for word alignment training from SC corpora, for each language pair.

3.2 Sense inventories

Thinking of word senses as discrete is challenging, because language is subject to change and interpretation by nature (Navigli 2009). Besides, the boundaries of each sense are arguable to tell: some dictionaries may make the decision to split two related and partially overlapping senses, while others would merge the two in a single entry.

A common strategy is to build sense inventories with an enumerative approach, i.e. by listing all senses for each word. Many have taken position against it, because of the impossibility in principle to fit the ever-changing semantic connotations of a word within a pre-specified number of senses (Kilgarriff 1997, 2006). However, in order to make possible a comparison of different WSD systems in computational tasks, a well-known and widely used sense inventory with enumerated senses is the way to go.

Generally speaking, a sense inventory is a computational lexicon or machine-readable dictionary (MRD) that describes the meaning of a word by listing all its senses. Ideally, such a dictionary should have full coverage and the sense distinctions should be clearly distinguishable from each other. Unsurprisingly, this is not always the case, since even humans disagree on what should be a sense; this is also why manually compiled sense inventories such as WordNet undergo revisions and have occasionally revised their internal structure; further details on WN are given more forward.

Another sense inventory came to be very popular in both monolingual and multilingual graph-based WSD research is BabelNet (Navigli and Ponzetto 2010), a multilingual knowledge base. While very large and constantly under development, BabelNet bootstraps word senses from very different sources, Wikipedia and WN itself on top of all, and in automatic way. On the other hand, WN and the like, although much poorer in coverage, can guarantee manual supervision on all the senses included in the lexical base, which may be preferable for some applications.

3.2.1 WordNet

WordNet (WN) (Fellbaum 1998; Miller and Fellbaum 1991; Miller et al. 1990) was created at Princeton University by George Miller's group and has been available for research purposes since 1993. Originally started for psycholinguistics studies on language acquisition in children, this project turned out to be the reference lexicon for WSD tasks in the NLP community. To this date, more than 40 projects have been launched with the goal of building wordnets in different languages.

WN is commonly considered one step beyond MRDs, since it encodes a rich semantic network of concepts. Since WN contains conceptualizations of specific domains and is organized in a taxonomy and includes a set of semantic relations, it can also be considered an ontology (Niles and Pease 2001).

Concretely, WN is an English lexical database in which nouns, verbs, adjectives and adverbs have each their own network and are linked to each other by means of various relations. In a network, nodes represent *concepts* and the edges connecting the nodes are *semantic relations*; the lemmas denoting concepts are also stored, and so are the *lexical relations* among them.

Everything is organized around the notion of *synset*: a synset is a **synonym set** that describes a sense by listing all the lemmas that may be used to express it, by giving a definition and usage examples and by listing all the relations with other synsets in the network. In order to determine that two lemmas express the same concept, synonymity does not have to be absolute: it is only required that one can substitute another in the same context without altering the truth value of the expression (Miller et al. 1990, p. 240).

It is important to highlight that concepts in WordNet must be lexicalised in the corresponding language, i.e. there must be at least one word that express it.

Semantic relations

Each network is organized around a main semantic relation. Nouns are structured according to **hypernymy**, which links more specific concepts (hyponyms) to more general ones (hypernyms) following an IS-A relationship. Since WN 2.1, there is a root node `entity` that subsumes all nouns below it. Before, there were 25 *unique beginners* picked to distinguish 25 fundamental semantic domains, now known as *supersenses*; this information is still carried in synsets' description, so it can be exploited to aid WSD.

WN 2.1 also drew a distinction between individual instances belonging to a type and subtypes of a broader type; in other words, it would be wrong to treat both the Nile and the concept `river` as types of `stream`, because the former is just an instance of a river and, more generally, of a stream, but it is not, technically, a *type* of stream.

Other relations can be registered for nouns, but they are optional: *meronymy* indicates that a concept is a part of another; following Winston et al. (1987), WN distinguishes between strict part meronymy, substance meronymy (the relation between a concept and the substance it is made of) and member meronymy (between a group and its members). See Fig. 3.3 for an example of the structure of the noun network in WN.

Verbs revolve around the relation of **troponymy**, although it is not recorded as consistently as for nouns and is actually encoded in terms of hypernymy and hyponymy. Troponymy links together synsets that are one a specific way of the

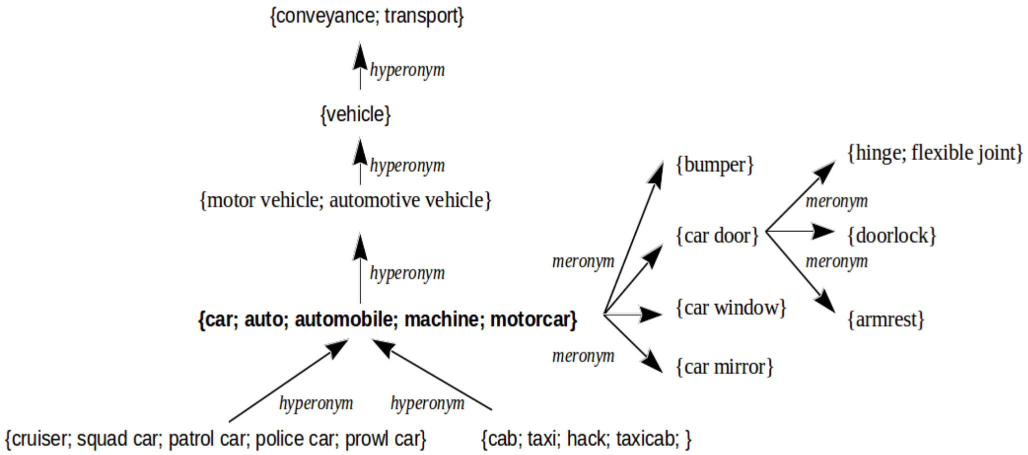


Figure 3.3: Portion of the noun network structure in WN 1.5. Readapted from Vossen (2002)

other, basically encoding the *manner* component; for instance, *to whisper* is a way of *talking*. As in the noun network, there are optionally more relations to record: *entailment* connects concepts in which one's meaning implies the other, such as in *live* and *exist*; *cause* expresses that two events are related because one causes the other, as in *kill* and *die*.

For adjective lemmas, there is a loose distinction in how relational and descriptive adjectives are stored. The former derive from nouns or verbs and are linked to them through the **pertainymy** relation. Instead, the latter express an attribute of a noun; this property makes it natural to organize descriptive adjectives around the notions of **antonymy** and **synonymity**. Antonymous lemmas are linked to each other by an antonym relation in the WN structure, and likewise synonymous lemmas are connected by a `similar_to` relation. In addition, adjectives are combined in clusters containing head synsets and satellite synsets, and

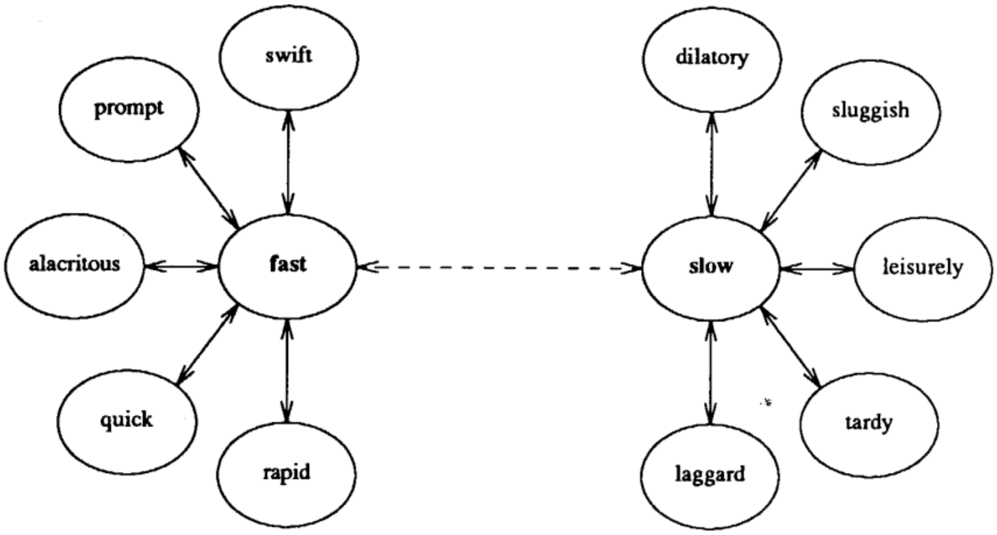


Figure 3.4: Example of the bipolar structure of descriptive adjectives in WordNet. From Fellbaum (1998, p. 51)

the head synset is connected, through the antonymy relation, to the head synset of another adjective cluster; see Fig. 3.4 and note that, while *quick* is certainly somehow opposed to *slow*, the most typical opposition a native English speaker would think of is *fast* ↔ *slow*, correctly represented through their head synsets.

Finally, adverbs are often derived from adjectives (as shown by pertainymy links) and, as such, they may have antonyms.

Fig. 3.5 (Morato et al. 2004) offer a nice overview of the range of applications for which WN has been used from 1994 to 2003; unsurprisingly, conceptual identification is by far the most natural task for which WN can come in aid.

WordNet deeply changed the approach to computational lexical semantics

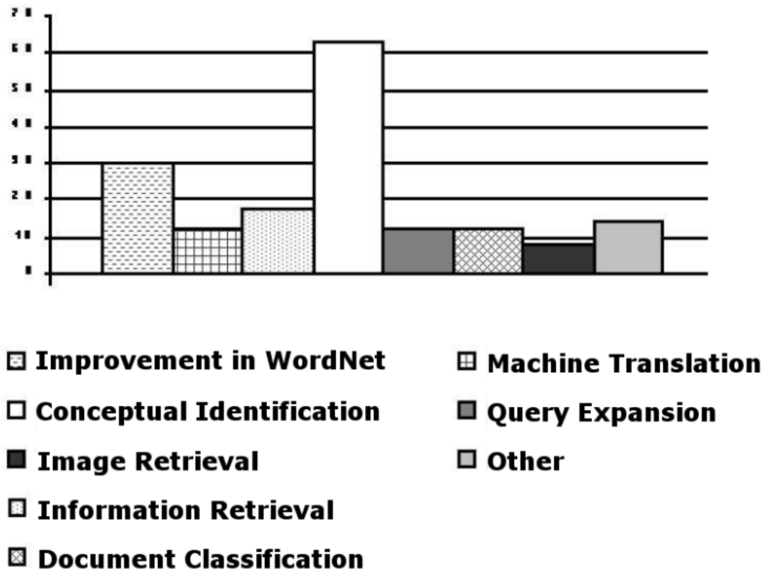


Figure 3.5: Range of WordNet applications in academic research from 1994 to 2003. Readapted from Morato et al. (2004, p. 261)

and was soon emulated in several other languages. The following sections describe in detail the WN projects that were employed to build translations of SemCor and that are, consequently, used also in this work.

3.2.2 MultiWordNet

MWN[10] (Pianta et al. 2002) is a multilingual lexical database strictly aligned with PWN 1.6: each Italian synset is linked to the corresponding English one. In fact, MWN was built following the so-called *expand model*, in which new wordnets are built starting from the synsets and the lexical and semantic relations in the English WN.

This approach is pretty straightforward and tends to guarantee higher interlingual compatibility. Pianta et al. (2002) point out that the building of any wordnet implies taking a series of decisions that can potentially increase the divergence with related WN projects, having this nothing to do with linguistic differences. The strategy of simply covering in another language what has already been done for another language takes this risk off the table.

Other WNs were produced to be compatible with the MWN model, but are not part of the MWN release; those are the Hebrew, Latin, Portuguese, Romanian and Spanish WNs, which can be all browsed through the MWN interface.[11].

[10] http://multiwordnet.fbk.eu/

[11] http://multiwordnet.fbk.eu/online/multiwordnet.php

MWN was created semi-automatically, in that lexical information was acquired relying on translation equivalents and an algorithm proceeds to attempt synset assignment. All data automatically acquired were manually checked. All in all, MWN includes information about 57,934 Italian word senses, 41,491 lemmas and 32,673 synsets (in correspondence with the English equivalents). 2,825 of the synsets were created from scratch and have no correspondent in PWN. In addition, 770 English-to-Italian lexical gaps were detected in the process. MWN maintains English lexical relations, but lacks the corresponding Italian ones, while semantic relations hold for both languages and 2,872 refer to the new synsets.

Besides lexical and semantic relations, MWN also represents correspondences between English and Italian concepts and semantic fields; this last project, known as WordNet Domains (Bentivogli et al. 2004), assigns every synset at least one domain label, choosing from a set of about two hundreds.

MWN is licensed under a Creative Commons Attribution 3.0 Unported License and is available both for research and commercial purposes. The currently available release is MultiWordNet 1.5.0.

3.2.3 ItalWordNet

MWN is not the only, nor the first, Italian WN. Roventini et al. (2000) built **ItalWordNet (IWN)** within the EuroWordNet project (Vossen 1998), a large project aiming at developing lexical resources for several European languages which took place from 1996 to 1999. During this time, WNs for Czech, Dutch, Estonian, French, German, Spanish and German were also produced.

EuroWordNet adopts a methodological framework distinct from MWN. Language specific WNs were built independently from each other, and only in a second phase was it attempted to find correspondences between them. This *merge model* requires more work, predictably, but allows for more freedom. In fact, English is not an interlingua, and attempting to exploit it as such leads to unreasonable constraints when trying to work multilingually, forcing the new wordnets to depend on the conceptual and lexical features of the source language. While this issue has been kept in mind all along and has now been recently addressed by the WN community (Bond et al. 2016), merge-derived WNs found their primary motivation in their property of addressing the peculiarities of languages in ways that better mirror the semantic properties, which may change across parts of speech.

In EuroWordNet lexical hierarchies of the wordnets involved are built independently from each other, and then aligned among them and to PWN in a second phase. They follow PWN in that synset is the fundamental unit and as for the semantic relations included. The lexical databases are interlinked via ILI, a language-independent, unstructured list of concepts. ILI served the purpose of representing a conceptualization of meanings that are, in turn, lexicalised by specific synsets in different languages. From a ILI record, multilingual conceptual search can be achieved without having to follow the English structure.

A major difference between EuroWordNet and PWN is that the former (and, consequently, IWN) does not adopt the traditional PoS distinction, but instead embraces the distinction among *semantic orders* made by Lyons (1977), which distinguishes: a) first-order entities: concrete things, perceivable through the senses and undoubtedly located in space and time, thus grouping together concrete nouns; b) second-order entities, being nouns, verbs and adjectives representing properties, states,acts, processes and events; c) third-order entities, which do not exist outside time and space and are conveyed by abstract nouns.

3.2.4 Romanian WordNet

BalkaNet (Stamou and Grigoriadou 2002) was a significant follow-up of EuroWordNet aimed to extend the multilingual lexical ontology with five Balkan area languages: Bulgarian, Greek, Romanian, Serbian and Turkish. Similarly to what happened to other sibling projects, the Research Institute for Artificial Intelligence of the Romanian Academy (RACAI) took RWN upon at the end of the BalkaNet project and has kept maintaining and developing it since then (Tufiş et al. 2008, 2013).

Within BalkaNet, the RWN had been mapped to PWN 2.0. RACAI research group developed its own mapping algorithm to port the annotations to PWN 3.0, and then validated it against the mapping automatically produced by the NLP Research Group at UPC,[12] scoring 95% precision. The non-matching instances were

[12] http://www.talp.upc.edu/index.php/technology/resources/multilingual-lexicons-and-machine-translation-resources/multilingual-lexicons/98-wordnet-mappings?highlight=WyJtYXBwaW5nIl0=

validated one by one, thus obtaining a quite reliable mapping.

RWN inherited from BalkaNet the so-called **BILI synsets**, synsets unique to Balkan cultures and languages. It later emerged that some of them double existing ILI synsets. Moreover, as later discussed, RSC annotation occasionally employs these tags, which lack a direct link to PWN despite being inserted in the hierarchy.

Mititelu et al. (2014) gives the latest information about the state of the resource: it contains 59,348 synsets, 85,238 words and 2,787 nonlexicalized synsets. RWN is licensed through META-SHARE and its use is free for academic research, but restricted for commercial use.

3.2.5 Japanese WordNet

The JWN (Bond et al. 2009; Isahara et al. 2008), originally developed by the National Institute of Information and Communications Technology (NICT) and firstly released in 2009, is a large-scale semantic dictionary of Japanese. The current release is version 1.1, available under the WordNet license. It contains 57,238 synsets (concepts) (all of which have definitions; 78% also show examples), 93,834 unique Japanese words and 158,058 senses (synset–word pairs).

The project started out as an attempt to quickly and efficiently build a Japanese version of PWN through the *expand* approach. In a first phase, English entries were translated into Japanese by exploiting a multilingual dictionary (Breen 2004) and other WNs that were already linked to PWN. Specifically, the German, French and

Spanish WNs were all mapped into WN 3.0 using the mappings from Daude et al. (2003).

For each synset in WN 3.0, the equivalents in the French, German and Spanish WNS are retrieved and, consequently, their Japanese translations. The Japanese equivalents are then ranked, with higher confidence scores given to the equivalents validated by more than one language. For 54.4% of the 117,007 WN 3.0 synsets a possible translation was found.

As for the manual development, the core synsets of WN and the synsets with the most frequent words were targeted first. In a second phase, the enlarging of JWN went along with the building of JSC, so the synsets occurring in MSC were translated first, along with 10,000 frequent words in the Juman dictionary.

From the releases following the first, increasing effort has been put in expanding JWN in ways that may also make it diverge from PWN, that is: a) by adding synsets unique to Japanese; b) when necessary, modifying the structure of the hierarchy so that it better represents the language. It has been estimated that 5% of the entries contain errors, which are meant to be fixed especially in the steps of translating the English glosses and sense tagging Japanese text.

Some errors will also be found and fixed by linking to other resources: the large public ontology SUMO (Niles and Pease 2001), the Japanese lexicon GoiTaikei and a collection of pictures from the Open Clip Art Library.[13]

[13] https://openclipart.org/

3.2.6 Open Multilingual WordNet

The Open Multilingual WordNet (OMW) is an open-source multilingual database that connects all the 34 open WNs linked to the PWN (Bond and Foster 2013; Bond and Paik 2012), for over $2,000,000+$ senses across $150+$ languages.

A convenient interface to OMW is provided by the Python module NLTK[14] (Bird and Loper 2004). It is possible to look up a lemma by using ISO-639 language codes.

OMW exists in a simplified (Bond and Paik 2012) and an extended version (Bond and Foster 2013), that includes further data extracted from Wiktionary and the Unicode Common Locale Data Repository. Costa and Bond (2015) provided a suite of web-based tools to expand and edit the WN projects included.

	Synsets	Senses	Words
English	117,659	206,978	148,730
Italian MWN	35,001	63,133	41,855
Italian IWN	15,563	24,135	19,221
Romanian	56,026	84,638	49,987
Japanese	57,184	158,069	91,964

Table 3.4: Coverage of the WNs used.

Table 3.4 gives basic coverage statistics for the WNs of our target languages. It

[14] http://www.nltk.org

should be noted that the figures refer to the actual numbers of synsets and word-sense pairs available after the merging process in OMW, as reported on the project website.[15] That is, the table reports the information that could actually be linked to and through PWN; every deviation from PWN (for instance, additional lexical and semantic relations) introduced by the original projects could not be maintained in OMW. Consider the case of IWN: taken alone, it has better coverage than MWN $(49,349$ vs $35,001$ synsets), but only $15,563$ of its synsets are mapped to PWN.

3.3 Requirements and preprocessing

This section outlines three important preprocessing steps that had to be carried out in order to be able to perform MSI.

The first two concern the only two requirements stated for MSI, that are the availability of a) a shared sense inventory; b) word alignment between any pair of the so-derived multilingual corpus.

The third preprocessing step anticipates an issue that will become apparent in the next chapter, i.e. cases in which the MSI algorithm is unable to make a decision based solely on WN information. In this case, heuristics based on sense frequency statistics can help to select the most likely sense among those found through MSI.

[15] http://compling.hss.ntu.edu.sg/omw/

3.3.1 Mapping to WN 3.0

In this work, MSI is performed over a multilingual corpus that consists of the texts that are covered in all four SC corpora. In order to build such corpus, the reference sense inventory must be the same for all parts. RSC and JSC are annotated with WN 3.0, but the English and Italian texts in MSC were annotated with WN 1.6, so their annotations have to be mapped to WN 3.0; this is, in fact, the most reasonable choice, because WN 3.0 is, at the present time, the most commonly used version (despite the fact that a more recent one, WN 3.1, has been available for years).

There are few more practical reasons: first, as it will be discussed in detail, the mapping between WN versions is reasonably precise; second, NLTK,[16] a suite of libraries available in Python for Natural language processing (NLP), provides a convenient interface to WN 3.0 and to OMW, making the look-up step in the MSI algorithm straightforward.

Conversion of the English annotations

English SC was annotated with *sense keys*, which reflect a subcategorization of word senses and are consistent across versions. It has been estimated that, on the sense keys alone, it is possible to correctly map around 95% of the WN 1.6 synsets to WN 3.0.[17]

[16] www.nltk.org

[17] According to the HyperDic project: http://www.hyperdic.net/en/doc/mapping

Actually, there is already a WN 3.0 version of SC: Rada Mihalcea from University of North Texas has made available SC releases annotated with different WN versions, WN 3.0 included.[18]

PoS	Annotations	Sense keys
Nouns	344	33
Verbs	336	37
Adjectives	136	7
Adverbs	1,021	5
Satellite adjectives	1,555	390
Total	**3,392**	**472**

Table 3.5: Distribution of lost sense keys in 116 English MSC texts

In Mihalcea's release of SC annotated with 3.0, all WN 1.6 synsets lost in WN 3.0 are maintained in the XML files, but are signaled with the tag `wnsn=0`; occasionally, the original annotation indicated more than one sense key, so it is often the case that at least one of the provided sense tags is still valid in WN 3.0. If only the 116 texts composing SC are considered, 3,392 out of 119,802 annotations (2.83%) have been lost. Table 3.5 shows how many annotations and sense keys have gone lost for each part of speech in WN.

The mapping from WN 1.6 to WN 3.0 has been carried out independently from Mihalcea's release. In the end, the two SC versions appear identical. Mihalcea's files needed some prepreprocessing: the XML was not valid because the values of

[18] https://web.eecs.umich.edu/~mihalcea/downloads.html#SC

the attributes were not in quotes; in a couple of cases there were empty attributes.

Since SC was manually annotated, there are occasionally multiple annotations for a single token. Consider for instance the following sentence pair in the text `r06`:

(EN) *"You and I have fallen out of literature into politics", Moreland observed.* (IT) *"Io e te siamo usciti fuori dalla letteratura e caduti nella politica", ha osservato Moreland.*

In SC the token "fallen" was annotated with both the synsets `spill.n.04` ("a sudden drop from an upright position") and `fall.n.05` ("a lapse into sin; a loss of innocence or of chastity"). In such cases, in MSC the first synset is consistently chosen for the projection to the Italian word, so the same criterion was applied for this mapping; in addition, also RSC and JSC have up to one annotation for token.

For the purpose of the task, the mapping was performed in order to maximize the number of the annotations. The mapping algorithm reads one SC file at the time and then loops over all the annotations, which are 119,802 for the 116 texts composing the English side of MSC. Of those, 116,410 are still valid in WN 3.0 (97%).

For each annotation, the retrieval of the WN 3.0 synset is attempted by looking up the sense key in the NLTK interface. This goes well in the vast majority of the cases (114,431/116,410, equal to 98.2%). 524 more annotations (0.45%) are found among the lemmas annotated with more than one sense key (up to three).

Finally, to maximize also the result of MSI, as a last resort the annotation of the aligned Italian word is looked up; if such annotation appears within the synsets associated to the English lemma in WN 3.0, then it is assigned to the English lemma. This happens in 1,545 cases, totaling up 1.33% more of the annotations maintained after the mapping.[19]

Conversion of the Italian annotations

Bentivogli and Pianta (2005) annotated the Italian side of MSC with WN 1.6, using an offset-based encoding. This encoding encapsulates an 8-digit number and the WN part of speech (`a`, `r`, `n`, `v`), respectively for adjective, adverb, noun and verb) and is extensively used in similar works, but it is not consistent across different WN versions. Fortunately, there are several freely available mappings between different WN versions[20] (Daudé et al. 2000; Daudé et al. 2001).

The mappings were automatically produced by exploiting both the graph structure and non-structural information (synset lemma names, glosses and verb frames). The error rate is assumed to be negligible; Daudé et al. (2001) report very good precision and recall scores for the automatic mapping of all parts of speech from WN 1.5 to WN 1.6 (97.9-99.3% and 99.9-100% respectively, depending on the part of speech).

[19] Two remarks should follow from this: first, English and Italian annotations for the same translation pairs may occasionally be different; second, at times the gold standard annotation will not be found between the synsets associated to the lemma in WN 3.0; those cases will be counted as separate during evaluation. See Chapter 4.

[20] http://www.talp.upc.edu/index.php/technology/tools/45-textual-processing-tools/98-wordnet-mappings/

PoS	Annotations	Sense keys
Nouns	15	6
Verbs	11	4
Adjectives	122	151
Adverbs	70	5
Total	**218**	**166**

Table 3.6: Distribution of lost sense keys in 116 Italian MSC texts

However, the changes occurred between WN versions 1.6 and 3.0 led to the loss of 218 annotations, as shown in Table 3.6. The corresponding tokens have been excluded from precision and coverage evaluation.

As mentioned in the previous section, the WN 3.0 Italian annotations were copied back on the English counterparts via the alignment whenever the sense key of the original English annotations was not valid in WN 3.0.

For the sake of clarity, Table 3.7 displays the number of annotations retained after the mapping for Italian and English texts in MSC; the table also reports statistics referred to the subset of texts that are shared with RSC and JSC as well. The difference in number of the source annotations is due to characteristics inherent to the Italian language and to individual choices of the annotators (Bentivogli and Pianta 2005).

Corpus	Annotations	English	Italian
49 texts	Valid	50,019	39,942
	Lost	1,201	60
	Total	51,220	40,002
116 texts	Valid	116,410	92,202
	Lost	3,392	218
	Total	119,802	92,420

Table 3.7: Retained annotations after the mapping to WN 3.0 for English and Italian MSC texts.

3.3.2 Aligning the multilingual parallel corpus

Word alignment is considered a hard NLP problem. Given two texts *T1* and *T2*, respectively in languages *L1* and *L2*, a word in *W1* in *T1* is aligned to a word *W2* in *T2* if the two words are reciprocal translations in their context, i.e. if they form a translation equivalence pair (Tufiş 2005). Other possible scenarios are the null alignments (when a word in *T1* has no counterpart in *T2*) and alignments in which multiple words are involved (*many to many* and *many to one* alignments).

Word alignment is not the only possible annotation between the two parts of a bitext. Alignments at the concept level, possibly spreading over more than one word, are also important and provide a different kind of information about the text that is complementary to the one given by the word alignment and, as such, worth to be made explicit (Bond et al. 2013).

1	Il		The	1
2	modo	04928903-n	way	2
3	in		she	3
4	cui			
5	ti			
6	ha			
7	parlato	00897564-v	addressed	4
8	mi		you	5
9	ha		made	6
10	fatto		me	7
11	perdere	01787822-v	lose	8
12	la		my	9
13	testa		mind	10
14	.		.	11

Table 3.8: Concept alignment

1	Il	The	1
2	modo	way	2
3	in	she	3
4	cui	addressed	4
5	ti	you	5
6	ha	made	6
7	parlato	me	7
8	mi	lose	8
9	ha	my	9
10	fatto	mind	10
11	perdere	.	11
12	la		
13	testa		
14	.		

Table 3.9: Word alignment

Tables 3.8 and 3.9 give an example of both alignments in an Italian-English sentence pair. As for the word alignment, there are examples of null alignments in both directions: Italian language needs *in cui* ”in which” for the sentence to be grammatical, while it can omit the subject *lei* ”she”, being a pro-drop language.

The concept alignment draws attention on the last block, which covers five words in both cases and assigns the concept `madden.v.01`: ”cause to go crazy; cause to lose one’s mind” to the idiomatic phrase *lose one’s mind*, which is identical in Italian. This example shows one of the main challenges of both the types of alignments, as for some people the phrase would be best represented as a single unit in preprocessing, just like what happens for multiwords. As Och (2002) express, very effectively, ”it is difficult for a human to judge which words in a given target string correspond to which words in its source string. Especially problematic is the alignment of words within idiomatic expressions, free translations, and missing function words. The problem is that the notion of correspondence between words is subjective.”

The following sections deal with exploiting both types of alignments to fulfill the requirement to perform MSI.

Aligning text with an automatic word aligner

The typical solution is to carry out automatic word alignment between every language pair. This option has the disadvantage that alignments are not verified by humans. A viable option would be to check a small sample of the automatic alignments produced and to estimate, from that, the precision of the overall alignment, but this requires a certain familiarity with the involved languages, which is not the case for Japanese.

Automatic word alignment needs as much training data as possible to be fed to the word aligner. The training data is to be sentence-aligned and word-aligned, so the system can learn from the given alignments.

For this project, the choice fell on **fast_align**, a word aligner that can produce alignments in absence of training data, being unsupervised.[21] fast_align is a log-linear reparameterization of IBM Model 2 that reportedly outperforms IBM Model 4, being at the same time ten times faster to train (Dyer et al. 2013).

The first step was to produce the raw versions of the SC texts from their embedding in XML files. The raw text extracted in this fashion is already tokenized, as Listing 3.3 shows for a sample sentence in all the MPC languages.

```
The Fulton_County_Grand_Jury said Friday an investigation of Atlanta 's
    recent primary_election produced " no evidence " that any
   irregularities took_place .

Venerdì il Grand_Jury_di_Fulton ha detto che un' indagine sulla recente
    elezione primaria di Atlanta non ha prodotto " nessuna prova " del
   fatto che si siano verificate delle irregolarità .

Marele Juriu din Fulton a spus vineri că o investigare a alegerilor
   recente nu a produs " nicio dovadă " că ar fi avut_loc nereguli .

フルトン 郡 グランドジュリー は 、 金曜日 に アトランタ の 最近 の 予備
    選挙 の 調査 で は 、 不正 行為 が 行なわ れ て いる か という こと
    に 、 「 徴候 なし 」 を 示し て い た 、 と 陳べ て い た 。
```

Listing 3.3: Examples of the same sentence from the raw texts.

Secondly, full-texts for each pair were produced. fast_align requires the input text to be tokenized and aligned into parallel sentences, where each line is a source language sentence and its target language translation, separated by a triple pipe symbol with leading and trailing white space (|||). Listing 3.4 shows an excerpt from the Italian-Romanian input to fast-align.

```
La sua istanza accusava di crudeltà psicologica . ||| Petiția lui acuza
    cruzimea mentală .
```

[21] https://github.com/clab/fast_align/blob/master/README.md

```
La coppia si_sposò il 2 agosto 1913 . ||| Cuplul s- a căsătorit pe 2
    august , 1913 .
Hanno un figlio , William_Berry_Jr . , e una figlia , la signorina_J._M
    ._Cheshire di Griffin . ||| Ei au un fiu , William_Berry_Jr. , și o
    fiică , d-ra_J._M._Cheshire din Griffin .
```

Listing 3.4: Excerpt from Italian-Romanian fast-align input data from SC text `a-01`

The aligner generates asymmetric alignments. The forward (source → target) and reverse (target → source) alignments can be symmetrized using a variety of standard symmetrization heuristics. The output is in the widely-used `i-j` Pharaoh format, where a pair `i-j` indicates that the `i`th word (zero-indexed) of the left language is aligned to the `j`th word of the right sentence. For example, the forward alignment of the above Italian-Romanian excerpt is shown in Listing 3.5.

```
2-0 3-2 5-3 6-4 7.5
1-0 2-1 2-2 2-3 3-4 5-5 6-6 7-8 8-9
0-1 1-2 2-3 3-4 4-5 6-6 7-7 8-8 9-9 10-10 11-11 12-12 13-13 14-14 15-15
```

Listing 3.5: fast_align output for the Italian-Romanian excerpt from SC text `a-01`

Once the forward and reverse alignments are produced, different symmetrization approaches can be attempted. Simple intersection of the forward and reverse alignment guarantees more precision to the cost of a smaller recall score, while the union of the asymmetric alignments gives the opposite result. Another common symmetrization approach is `grow-diag-final`, which starts with the intersection of the two alignments and then adds additional alignment points in their neighborhood. Once the alignments have been produced, a python script loads and looks them up for every token to be disambiguated. In Chapter 4 MSI is performed comparing the results obtained using first the alignments produced by intersection and then those obtained through grow-diag-final.

Additional training data for Italian-Romanian - A good resource for EU languages is the DGT-Translation memory (Steinberger et al. 2013). This collection gathers the texts in the multilingual legislative documents of the European Union (Acquis Communautaire"), which were released by the European Commission in the effort of supporting multilingualism.

The resource consists of aligned corpora for 276 language pairs from the 24 EU languages. The translations are manually produced and the distribution comes with a software that can produce any bilingual parallel corpus from the texts selected.

The Italian-Romanian parallel corpus that can be produced from this files consists of 442,814 aligned sentences, which need tokenization before being passed as input to fast_align.

Additional training data for English-Romanian - In addition to the 82 RSC texts, more data for training fast_align for English-Romanian can be found in Rada Mihalcea's website:[22]

- a Romanian-English dictionary (38,000 entries).
- 1-million-word Romanian-English parallel texts. This collection groups together the parallel text of Orwell's novel "1984", the Romanian Constitution, and a large (about 900,000 tokens) collection of newspaper texts collected

[22] https://web.eecs.umich.edu/~mihalcea/downloads.html#romanian

from the Web, to a total of 50,248 parallel sentences. This was originally training data for the Romanian-English word alignment task held for the HLT-NAACL 2003 workshop. [23] All texts are sentence-aligned and tokenized.

- Hand-checked word aligmments for 17 English-Romanian sentences, plus 248 parallel sentences which constituted the test data for the task in the above workshop.

- 469,971 aligned sentences from the DGT-Translation memory (see above), which had to be tokenized first.

Additional training data for English-Japanese - Utiyama and Isahara (2003) released parallel data for Japanese-English, which include 150,000 1:1 sentence alignments and 30,000 1:many sentence alignments in the version for research purposes.[24]. The sample available on the website includes 100 sentence pairs.

A much larger resource is the Japanese-English Bilingual Corpus of Wikipedia's Kyoto Articles, which consists of 487,230 sentence pairs.[25]

In both cases the text is raw, i.e. it needs to be preprocessed and tokenized.

[23] http://web.eecs.umich.edu/~mihalcea/wpt/

[24] http://www2.nict.go.jp/univ-com/multi_trans/member/mutiyama/jea/index.html

[25] http://alaginrc.nict.go.jp/WikiCorpus/index_E.html

Aligning text through sense

A different approach consists in exploiting the sense annotations available in each SC corpus and try to align every side to its counterparts using matching annotations as hints. This has the advantage of resulting as valid as human-checked; however, this method is strictly dependent on the quality of the mapping, as different SC corpora were originally annotated with different WN versions, and the quantity of actual annotations per sentence (drastically fewer both in RSC and JSC). Conveniently, the sentence alignment is available for all pairs, because it mirrors the sentence splitting in the original English SC.

For each aligned sentence pair, the algorithm loops over the tokens, one language at the time. Every content word pair `<source_word,target_word>` is a candidate pair in the alignment task. First, all candidate pairs sharing the same sense annotation are paired together. If any words are left unaligned after this step, the remaining alignments are inferred by taking into account PoS information and synset similarity scores.

Let us imagine that the English and Romanian texts are to be aligned. Suppose the first step alone has aligned all Romanian content words but one, and that the corresponding English sentence has three content words left that are candidates for the alignment. Then, the aligner computes the most likely match by looking for PoS correspondence and for higher proximity in the WN network, by looking at a combination of the *path similarity score* and the *shortest path distance*.

This latter alignment strategy (the only possible source of errors) achieved

97% precision on a small sample (12%) of the alignments found for the language pair English-Romanian. It should be pointed out that only the second step might introduce errors, because heuristics come into play to align the remaining unaligned content words.

3.3.3 Sense frequency statistics

MSI employs sense frequency statistics (SFS), if available, to make a decision in case the comparison with other languages did not clear all the ambiguity. As often reminded in this work, sense-annotated data are scarce, and so is, consequently, the availability of sense frequency data.

For this reason, MSI also employs statistics extracted from the SC corpora themselves. However, in order to avoid a positive bias, the disambiguation step takes one text at the time as input and, when resorting to SFS, it excludes the frequency data computed over that text.

The frequency data have been extracted and stored in json files, which are loaded during the disambiguation step of MSI. Listing 3.6 shows an example of accessing the dictionary created for Italian for the lemma *signore* ”mister”; the WN 3.0 offsets found in MSC are reported with the respective frequencies in all the texts in which they occur. In text `p07` the offset `10601451-n`, corresponding to synset `sir.n.01`: *term of address for a man*.

```
>>>dict_ita['signore']
{'06341340-n': {'n05': 2}, '10388440-n': {'k10': 1}, '09536363-n': {'
```

```
    f10': 1, 'k19': 4}, '10127273-n': {'l12': 1}, '10601451-n': {'n12':
    1, 'k21': 3, 'p07': 4}}

>>>dict_ita['signore']['10601451-n']['p07']
4
```

Listing 3.6: Example of the dictionary storing sense frequency data from SC for Italian.

In addition, there are more resources from which it is possible to extract sense frequency data for English. Rada Mihalcea made available the Senseval-2 and Senseval-3 English all-words data (Edmonds and Cotton 2001; Mihalcea and Edmonds 2004) in SemCor format.[26].

Most of the SFS used in this work, however, come from the WordNet Gloss Corpus, from which 506,805 annotations were extracted. The corpus contains WN synset glosses that are, in turn, annotated.

```
<synset id="n00003553" ofs="00003553" pos="n">
 <terms>
  <term>whole</term>
  <term>unit</term>
 </terms>
 <keys>
  <sk>whole%1:03:00::</sk>
  <sk>unit%1:03:00::</sk>
 </keys>
 <gloss desc="orig">
  <orig>an !\colorbox{lightgray}{assemblage}! of parts that is
    regarded as a single entity; "how big
is that part compared to the whole?"; "the team is a unit"</orig>
 </gloss>
[...]
 <gloss desc="wsd">
[...]
 <id id="n00003553_id.4" lemma="assemblage" sk="1:14:01::" />assemblage
    </wf>
[...]
 <id coll="a" id="n00003553_id.5"
```

[26] https://web.eecs.umich.edu/~mihalcea/downloads.html#wa

```
    lemma="regard as" sk="regard_as%2:31:00::"/>
[...]
33  <id id="n00003553_id.1" lemma="entity" sk="entity%1:03:00::"/>entity</
    wf>
[...]
36   <ex id="n00003553_ex1">
[...]
47   <id id="n00003553_id.2" lemma="whole" sk=whole%1:03:00::"/>whole</wf>
[...]
51   </ex>
52   <ex id="n00003553_ex2">
[...]
59   <id id="n00003553_id.3" lemma="unit" sk="unit%1:03:00::"/>unit</wf>
[...]
62   </ex>
63  </gloss>
64 </synset>
```

Listing 3.7: Extract of the WN Gloss Corpus. Sense frequency data are extracted from the annotations present in glosses and usage examples (highlighted).

Annotations were extracted from the merged files of the release, from both glosses and usage examples. Listing 3.7 shows an extract of the WN Gloss Corpus. For each `synset`, its `terms` and respective sense `keys` are listed. Lines 10-12 contain the original `gloss` text and the annotations start on line 16, first going through the gloss words (lines 16-35) and then the usage examples (lines 36-51 and 52-62). Within example sentences, only the synset lemmas are disambiguated.

The lines highlighted in Listing 3.7 are the extracted lemma-sense pairs, while the annotations referring to the synset lemmas are ignored. The pairs are sorted, counted and stored in a dictionary, that is consulted whenever the algorithm is not able to return a single annotation.

Discontiguous spans of text have special markup; for instance, the phrase

”personal or business relationship” becomes `personal_relationship` and `business_relationship`, while phrasal verbs are reunited: ”pay them off” becomes `pay_off`.

At least for English, the lack of unbiased sense frequency statistics can be then addressed by including the above, addressing for 157,300 lemma-pos pairs.

In the following chapter all of these ”behind-the-scenes” requirements will find their place in the MSI procedure.

Chapter 4

Implementation and evaluation

This chapter describes in detail the implementation of MSI and discusses the results achieved with a multilingual corpus of 4 languages: English, Italian, Romanian and Japanese.

The theoretical grounds behind MSI is in that an ambiguous word will often be translated in different words in another language. A polysemous word in the target text can be easily disambiguated if its translation in another language is monosemous. As a consequence, the knowledge of all the senses associated to its translation can help detect the sense actually intended in the original text.

However, especially in closely related languages, polysemous words may convey the same senses and hence end up holding exactly the same semantic ambiguity. Consider, for instance, the English word ”interest” and its Italian translation ”interesse”, which share as many as four senses in WN 3.0:

```
interest.n.01 - a sense of concern with and curiosity about someone or
    something
interest.n.03 - the power of attracting or holding one's attention (
    because it is unusual or exciting etc
interest.n.04 - a fixed charge for borrowing money; usually a
    percentage of the amount borrowed
interest.n.05 - (law) a right or legal share of something; a financial
    involvement with something
```

Listing 4.1: Example of parallel ambiguity.

Nonetheless, the more the languages available for comparison in the parallel corpus, the more likely is that MSI actually manages to discern the correct sense in context. Consider, for instance, the problem of disambiguating the English word *administration* in Example 1.

(1)(EN) *The jury praised the administration and operation of the Atlanta Police Department.*

(IT) *Il jury ha elogiato l'amministrazione e l'operato del Dipartimento di Polizia di Atlanta.*

(RO) *Juriul a lăudat administrarea şi conducerea Secţ iei de poliţie din Atlanta.*

(JP) 陪審団は、アトランタ警察署の 陣営 と働きを賞賛した。

Given the alignments, we can retrieve the set of synsets associated with the lemmas in the Italian, Romanian and Japanese translations. Figure 4.1 shows how the intersection helps detecting the correct sense, which is the only one shared by all the lemmas.

Most often, however, such a comparison will only partially reduce the ambiguity, especially as such a fine-grained sense inventory as WN is used. Yet, other approaches (employment of human annotators, or recourse to baselines) can be

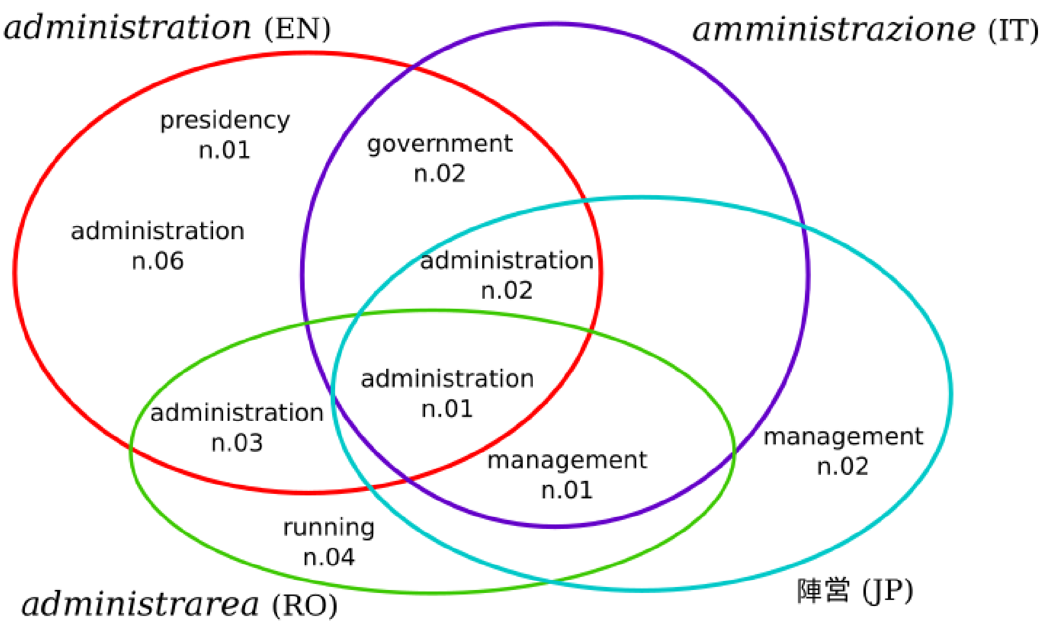


Figure 4.1: Disambiguation via MSI

applied in a second phase to solve the disambiguation task, once it has been simplified.

Differently from previous work (Bonansinga and Bond 2016; Bond and Bonansinga 2015), the present one focuses on the subset of the corpus shared across all four components and for which there are alignments. Unfortunately, this also means testing MSI on an even smaller corpus, but hopefully this way the contribution brought by three more languages can become more apparent. Moreover, unlike SP, MSI does not require any of the texts in a parallel corpus to be sense-annotated, so it can be applied to a wider range of existing resources.

4.1 MSI algorithm

```
for each language corpus in MSI
  for each word to be sense-annotated
    look for aligned words in the parallel corpus
        given this translation tuple, assign each word its set of
           synsets in WordNet
        perform intersection progressively over each non-empty set of
           senses that has been retrieved
        end when the overlap contains one sense, else make a decision
           using heuristics
```

Listing 4.2: MSI algorithm outline.

```
if |overlap| = 1 → Disambiguated
else if |overlap| > 1
  if available, intersect with general sense frequency statistics for
     the target lemma
    then:
      if |overlap| > 1, select sense with the relative highest
         frequency → RMSF_within_overlap
      else if |overlap| = 1 → MSF_in_overlap
      else, assign MFS → MSF
   else:
       resort to WN first sense (for English) or available SFS
       if MFS is found:
           if |overlap ∩ mfs|, select MFS → MFS_in_overlap
           else → MFS
       else:
           select random sense in overlap → random_in_overlap
```

Listing 4.3: Possible scenarios once the overlap is computed.

The algorithm disambiguates one side of our multilingual parallel corpus at a time, having as target the aligned texts in the other languages, when available. The intersection step proceeds by intersecting the target language synset set with the others, one at the time, so to avoid intersections that lead to an empty set. Listing 4.2 outlines the basic functioning.

If the *overlap* only consists of one sense, then the target word is `Disambiguated`.

If the overlap contains more than one candidate sense and there are SFS at disposal, then the overlap is further intersected with the set of most frequent senses available for the target lemma.

Listing 4.3 shows the possible outcomes that may emerge. If resorting to SFS leads to an overlap containing one sense, the word is disambiguated (`MFS_in_overlap`); if the overlap still results in more than one sense, the most frequent one among the ones left is selected (`RMFS_within_overlap`). If no other language contributes to disambiguate, we assign the current target lemma its MFS (`MFS`).

In case SFS are not available, a decision is made anyway by picking randomly from the overlap (`random_in_overlap`). Of course, MSI cannot be performed if the target lemma has no longer a sense in WN 3.0 or in lack of word alignments.

Concretely, the software first loads the multilingual corpus. Accepted input formats are JSON files or a self-contained XML file compliant to the NTUMC DTD (see Chapter 5). The multilingual corpus includes, for each component, at least lemma and part of speech information. Gold-standard sense annotation, if available, can be used for later evaluation. Alignments at all levels - document, sentence, word - need to be included too.

OMW is accessed through NLTK interface and MSI works with all languages present in it.[1]

[1] RWN has not always been part of the OMW, so in previous work it used to be loaded as an external resource for lemma and synset lookup.

4.2 Evaluation

4.2.1 Evaluation metrics

Gale et al. (1992a) posed the problem of a meaningful evaluation and firstly proposed upper and lower bounds against which the performance of disambiguation algorithms could be compared. A traditional **upper bound** metric consists in measuring the human inter-annotator agreement (IAA) on the same data (Gale et al. 1992a), as the expectations are that different systems will not be as consistent as humans. The reasoning behind this choice is to find a way to express that a text is too hard to disambiguate, even for human judges, thus the low IAA (Tufiş 2006).

On the other hand, the simplest baseline is choosing the **most frequent sense** (MFS). A baseline should represent an expected lower bound on the performance of automatic systems and normally indicates whether "a more complicated system is worth the additional implementation effort" (Palmer et al. 2006, p79). In very early work, Gale et al. (1992a) estimated lower and upper bound for a WSD, two-way ambiguities to be 75% and 96.8%, respectively; the MFS baseline is indeed very hard to beat; in contrast, the upper bound score crucially depends on the annotator's choices, the sense inventory available, etc.

Palmer et al. (2006) and Navigli (2009) discuss the most common scoring systems for WSD and related problems. The simplest one is to assign a score of 1 for each correct sense tag, and 0 otherwise, where a correct sense tag is one that matches the one assigned by the human annotator(s).

A system is evaluated in terms of coverage, precision and recall. **Coverage** is the percentage of words for which the system guesses the correct sense tag. **Precision** is obtained by dividing the number of guessed words for the number of guessed-on words. **Recall** is computed by dividing the number of answers provided (counting unguessed-on items as zero score) for the total number of items in the evaluation set.

4.2.2 Coarse-grained evaluation

> "Finer sense distinctions are only relevant as far as they get lexicalized in different translations of the word". (Lefever and Hoste, 2014)

According to Ide and Wilks (2006), coarse-grained sense distinctions are the only ones that we can consistently and coherently discern between. Research shows how even human annotators can have trouble agreeing upon the correct sense if sense distinctions are too detailed and specific. Ide and Wilks (2006) argue that the customary fine-grained division of senses pursued by lexicographers is not what we should aim for the computational task of WSD. Specifically, the fundamental distinction needed for NLP corresponds roughly to being able to discern homographs or etymologically related senses that are 'distinct for humans as homographic ones', *and nothing beyond that*.

Motivation for this strong, although well-founded, claim comes from the need to actually assist NLP applications with high-quality WSD modules, as they could hardly benefit from a mediocre performance. Their suggestion hence is that WSD should be performed at a level where optimal results can be reached, namely the one that mostly tries to distinguish homographic sense distinctions.

Sense inventories are a crucial part of this approach. Not only are a sufficient coverage and the alignment to the Princeton WN necessary: when it comes to deciding how to define close, very specific senses, a trade-off between the detail of the sense description and its actual usability in real contexts is highly desirable.

The fine granularity of WN senses can occasionally, depending on the application, be more of a practical disadvantage than a quality. In this analysis, for instance, error analysis suggested that the senses found through MSI were often very close, but it may happen that they are discarded as wrong outputs just because one language has a WN more developed and granular than another. We should also bear in mind that the correct senses against which we evaluate were picked by trained human annotators in the first place, and human annotators tend to describe a word as precisely as possible.

Conscious of this limit, Navigli (2006) devised an automatic methodology to find a reasonable sense clustering for the senses in WN 2.1. Sense clustering can be of great help in tasks where minor sense distinctions can be ignored, allowing a coarse-grained evaluation. They found 29,974 main clusters, some of which were manually validated by an expert lexicographer for the Semeval all-word task.

The original procedure requires a license to the Oxford Dictionary of English, which Navigli (2006) exploited to obtain a mapping to coarser senses. However, the clusters found can be easily converted to WN 3.0, as sense keys were used to point to synsets. Hence, we mapped the senses in the clusters found to WN 3.0, losing 101 of them in the process (typically one-element clusters); see section 4.2.3 for the discussion of our coarse-grained evaluation of MFS baseline and MSI.

At the fine-grained level, only identical sense tags count as a match, while at the coarse-grained level, all sense tags given in the gold standard and system guesses are mapped to the top-level sense tag, and the system receives a score of 1 if its guess has the same top-level sense as the correct tag. (Palmer et al. 2006)

4.2.3 Current results

Table 4.1 shows the precision and recall scores achieved with MSI and, for comparison purposes, the MFS baseline.

MFS performs very well for English; this is certainly due to the frequency data derived from SC, which produce a bias. We remind that WN senses are ranked with respect to the frequency counts computed on the semantic annotated parts of the Brown Corpus, i.e the larger set of the MPC hereby disambiguated. Thus, MFS should supposedly perform not as well for other parallel corpora, making MSI a viable and inexpensive cross-lingual disambiguation approach.

Generally speaking, it can be seen that the contribution of four languages is decisive to reduce ambiguity and make a better decision, even when based on SFS. The very high results in Romanian and Japanese, although encouraging, need to be read while considering that they refer to the content words for which we had alignment to the English SemCor (and, consequently, manual annotation to check against); the annotations in these two SemCor projects are significantly fewer than in the English and Italian counterparts (see Table 3.2).

In Table 4.1 we also show the improvement in precision obtained thanks to coarse-grained evaluation, which is significant for all languages. This could suggest that the results are especially corpus-dependent, as the manually assigned correct senses against which we evaluate are very specific. In fact, the senses found by intersection would be just good enough in most cases.

Of course, coarse-grained evaluation causes the MFS baseline to improve as well. In the case of English - which, again, is the component most subjected to the bias introduces by SFS - coarse-grained MFS still performs better than coarse-

Method	English		Italian		Romanian		Japanese	
	P	R	P	R	P	R	P	R
MFS (baseline)	**0.759**	**0.998**	0.623	**0.999**	0.723	**1**	0.831	**1**
MSI	0.692	0.712	**0.691**	0.966	**0.749**	0.728	**0.874**	0.731
Coarse-grained MFS	**0.837**	**0.998**	0.698	**0.999**	0.806	**1**	0.878	**1**
Coarse-grained MSI	0.793	0.875	**0.766**	0.966	**0.837**	0.728	**0.918**	0.712

Table 4.1: Precision (P) and Recall (R) scores obtained by MSI and MFS baseline on MPC (subset of 49 texts), compared with the respective scores with coarse-grained evaluation.

grained MSI.

All the other languages show higher scores for coarse-grained MSI. For Romanian, the improvement achieved through coarse senses is the most striking, while for Japanese we obtain the highest score overall.

Compared to the results discussed in Bonansinga and Bond (2016), in this analysis all languages give their contribution in the disambiguation process for the others. A substantial difference is also in the employment of additional word alignments and sense frequency statistics external to SemCor.

4.2.4 Overall contribution of sense intersection to MSI

Table 4.2 shows the distribution of all of the possible scenarios that may emerge when applying MSI, with reference to the possible outcomes in Listing 4.3.

The different scenarios are reported only for words correctly disambiguated. Generally speaking, in all languages MSI by itself manages to find the correct sense in half the cases and, as maybe expected, works particularly well for Japanese, which is the language farthest away from all the others.

The next scenarios through which MSI comes to a terminating condition is by employing SFS. When the assignment type is `MFS`, it means that the correct sense actually got lost through multiple overlaps; luckily, this happens very rarely.

Finally, as mentioned before, WordNets in languages other than English are not as rich, so it is often the case that no sense is found or, more likely, the target lemma name is missing in the synonym set, even though the synset is available in the foreign WordNet under other translations.

%	EN	IT	RO	JP
Disambiguated_by_MSI	61.90	51.73	47.87	66.57
MFS_in_overlap	3.36	13.07	12.29	12.76
RMFS_within_overlap	34.63	31.64	36.87	11.49
MFS	0.10	0.02	0.19	0.20
Random_in_overlap	0.00	0.09	0.13	0.01
No sense found in WN 3.0	0.01	3.45	2.65	8.96

Table 4.2: Distribution of Coarse-MSI outcomes.

We also compute **Average Ambiguity Reduction (AAR)** scores for all languages; see Table 4.3. AAR is calculated by comparing the number of all possible senses for a given lemma in WN with the number of senses left in the overlap after sense intersection with any aligned words available in other languages.

Component	AAR
English	0.43
Italian	0.50
Romanian	0.62
Japanese	0.50

Table 4.3: Average ambiguity reduction for all components of MPC.

4.3 Error analysis

4.3.1 Addressing unsuccessful intersections

In some cases, the correct sense is just not available in the other languages, so MSI discards it by progressive intersections. This is a problem inherently due to the coverage of the WNs involved. An obvious countermeasure is to share exactly this kind of outcome with the maintainer of the respective WNs, who would have the benefit of knowing exactly what senses and lemmas are the more urgent to add to their resources along with a helpful list of corresponding synsets and definitions across other languages. In the best scenario, with full sense overlap, the task of enriching WNs could become as straightforward as connect the suggested missing synset as it is connected in, say, English, with the only actual trouble of translating the definition.

Table 4.4 lists a few examples of cases in which MSI selects a sense that does not correspond to the one assigned by the annotators, but could be acceptable in context. However, cases like this will score 0 because of the non-match. Again, a coarse grained evaluation could come in help in addressing this issue; see Section 4.2.2.

Correct sense	Selected sense	Disambiguation
late.s.03	recent.s.01	MFS (in overlap)
produce.v.04	produce.v.01	MFS (in overlap)
evidence.n.01	evidence.n.02	Disambiguated by MSI
abnormality.n.04	irregularity.n.02	Disambiguated by MSI
end.n.02	end.n.01	MFS (in overlap)
own.v.01	have.v.01	MFS (in overlap)
war.n.02	war.n.01	MFS (in overlap)
drop.v.02	drop.v.01	Disambiguated by MSI
play.v.06	play.v.03	(relative) MFS (in overlap)
pure.s.04	pure.a.01	Disambiguated by MSI
unlock.v.03	unlock.v.01	Disambiguated by MSI
kernel.n.03	heart.n.01	MFS (in overlap)
forever.r.02	everlastingly.r.01	Disambiguated by MSI
meaning.n.02	meaning.n.01	MFS (in overlap)
matter.n.01	matter.n.03	Disambiguated by MSI
delay.v.01	delay.v.02	Disambiguated by MSI
apparent_motion.n.01	motion.n.03	MFS (in overlap)
coalesce.v.02	blend.v.03	MFS (in overlap)
operate.v.03	control.v.01	MFS (in overlap)
information.n.05	randomness.n.01	Disambiguated by MSI
earth.n.01	universe.n.01	MFS (in overlap)
match.v.01	fit.v.04	(relative) MFS (in overlap)

Table 4.4: Examples of senses found by the algorithm vs corrected senses

4.3.2 Limitations

MSI presents some known issues, that are due to both external and inherent causes. The large variety in the coverage of WNs belongs to the first group; MSI, and other knowledge-based WSD systems likewise, crucially relies on the quality of the WN projects involved. This just states the point, one more time, that there is a concrete need for development of lexical resources to aid WSD. In Appendix B we report the lemmas found most missing across the involved WordNet projects along with the suggested WN 3.0 synset offset, so to provide the maintainers with useful feedback for future work.

In addition to issues due to external causes, the task is difficult in itself even for human annotators, as inferable from the IAAs reported for SC and Italian MSC, respectively 78.6% and 81.9%. Finally, sense granularity constitutes another issue that would be independent from MSI but, unlike the first two, it can be addressed by using coarser senses.

In contrast, an inherent limit of MSI is that it needs SFS in order to make a decision in cases of uncertainty; SFS are hard to come by, so the lack of them has definitely an impact on the algorithm performance. Theoretically, the more the languages in a multilingual parallel corpus, the fewer the cases of uncertainty, but this is true, again, as long as the respective WNs are rich in coverage and properly inter-linked with other projects based on PWN.

The need for SFS could actually be cut if MSI was used to just annotate what it is able to, without resorting to SFS to be able of making a decision in all cases. This way, the output of MSI could be used as training data for WSD algorithms that do not require fully annotated data (like UKB, see Agirre and Soroa (2009)).

Chapter 5

Towards reproducibility

This chapter is about making the work of this thesis available to a greater audience. The ultimate goal is to make the suggested approach to MSI easily reproducible: hence, the code developed for this project is being made freely accessible at `https://github.com/jusing-es/clwsd`.

MSI works out-of-the-box with corpora that respect the format defined by the official NTUMC DTD (see Section 5.1).

Furthermore, the tool produces output in a format that is accepted by the pipeline that imports new corpora in the NTUMC database. The integration with the NTUMC tools is further discussed in Section 5.2.

5.1 From a portable XML to NTUMC and back

As we think of a format that would enhance any operation from/to the NTUMC database, it is useful to have the current NTUMC structure in mind.

5.1.1 The current NTUMC schema

NTUMC comes with a handy interface[1] that allows searching over the multilingual corpus for each kind of unit (tokens, lemmas, parts of speech, concepts).

Being built as a `sqlite3` file, the monolingual database accepts data in `.csv` format by default. The core tables in the schema and the way they depend on each other are represented in Figure 5.1.

Each new corpus is assigned a unique id in the `corpus` table. Table `doc` is used to register information specific to the single text unit, such as title and language.

Sentence segmentation and tokenization are the minimum requirements for any corpus to be imported. Table `sent` is to be populated with as many rows as there are sentences; the original raw text is always recorded as well. As it follows, all tokens populate table `word` with surface form, lemma and PoS. Each of these units is assigned a unique id that allows cross-references.

Finally, all the sense-annotated lemmas in the database populate the `concept` table: each annotated word is assigned a concept id and registered with its lemma

[1] http://compling.hss.ntu.edu.sg/ntumc/

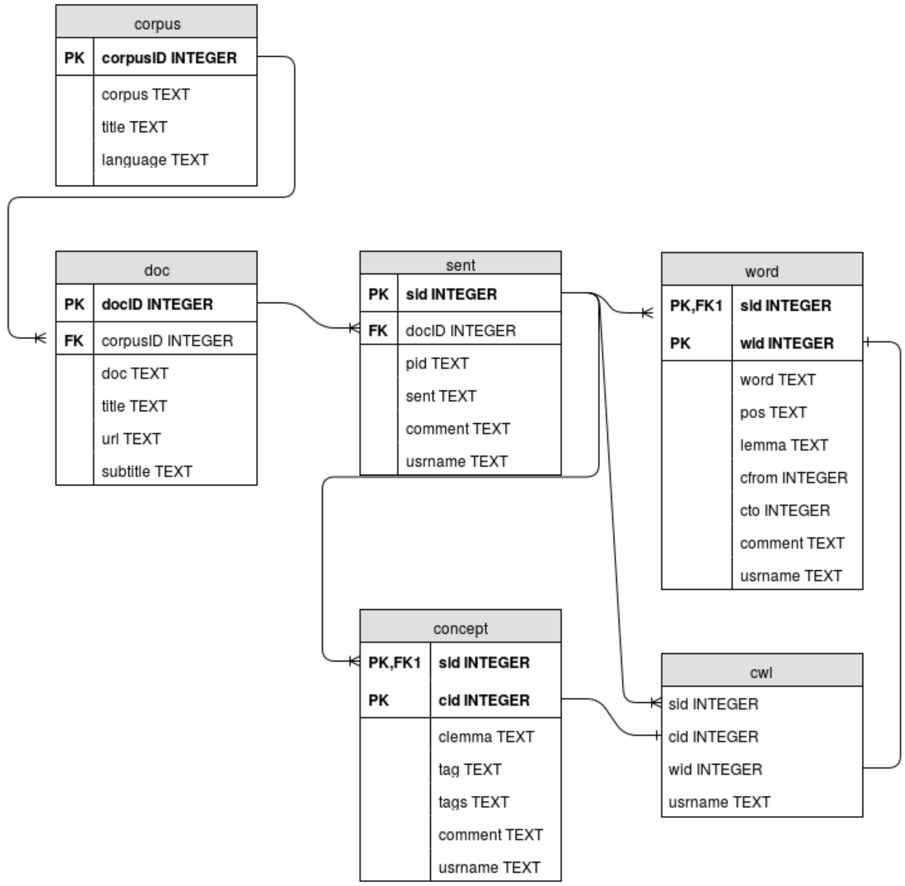
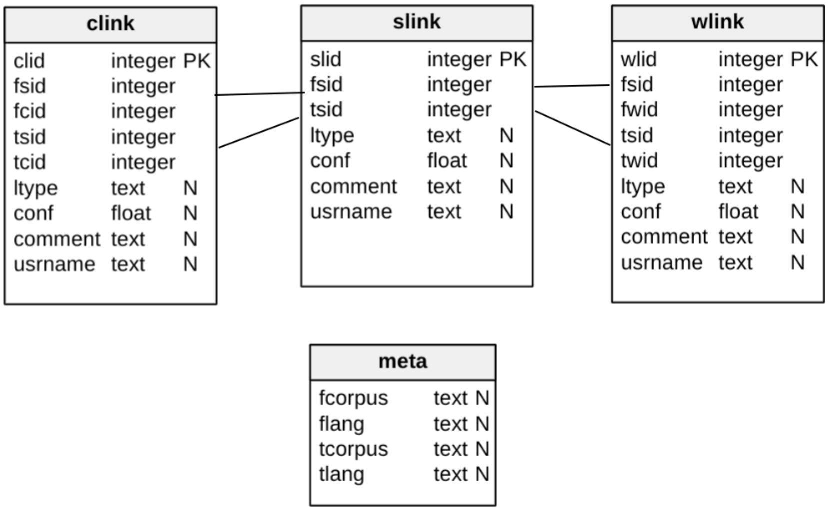


Figure 5.1: Schema of the NTUMC database.

and sense tag, in the form of WN 3.0 offsets. Table `cwl` functions as bridge between `concept` and `word` tables, providing the link between the respective ids. The link to the corpus is guaranteed by the sentence id.

In case the corpus to be imported comes already with any levels of annotations, the original ids are also registered for future reference, even though, within the database, all units discussed above must take a new id. If the corpus is multilingual, then a database for each language (if not existing) and a database for the language pair is created, which serves to store the alignment information at the sentence, word and concept level, as in Figure 5.2.



clink

clid	integer	PK
fsid	integer	
fcid	integer	
tsid	integer	
tcid	integer	
ltype	text	N
conf	float	N
comment	text	N
usrname	text	N

slink

slid	integer	PK
fsid	integer	
tsid	integer	
ltype	text	N
conf	float	N
comment	text	N
usrname	text	N

wlink

wlid	integer	PK
fsid	integer	
fwid	integer	
tsid	integer	
twid	integer	
ltype	text	N
conf	float	N
comment	text	N
usrname	text	N

meta

fcorpus	text	N
flang	text	N
tcorpus	text	N
tlang	text	N

Figure 5.2: Schema of the NTUMC database for a language pair.

5.1.2 Designing a standard XML input format

While the default import via CSV files is handy, NTUMC is growing and it would be ideal to make this process as simple and standardized as possible. This can be achieved by designing a customized XML format that satisfies all of the requirements in the database. As part of this enhancement, the current pipeline should be enriched with a new tool that will parse any corpus in this XML format and will insert it automatically into the NTUMC database.

Contributions from other projects

The scientific community has always made an effort to establish an XML standard to tag text, as the inherently hierarchical markup comes handy when describing several annotation layers. Moreover, XML encodes text in UTF-8 by default, which is a good practice.

With a standard XML template, many recurring issues would be taken care of in a consistent and robust way. Corpora are built hierarchically and all of their elements need to reference each other: that is, requirements such as the uniqueness of the ids and the clear relationship between related units (such a sentence and its tokens) is offered out-of-the-box by XML. Moreover, corpora may be multilingual and have alignments at each level - document, sentence, word, concept - and it might be challenging to come up with a simpler format that successfully manages all of these complexities.

In this regard, two formats at least deserve to be cited, as the good principles thereby applied served as inspiration to the NTUMC's own template.

The **Corpus Workbench XML** (CWB) is a collection of open-source tools built to query large corpora enriched with multiple layers of word-level annotation. Its powerful Corpus Query Processor (CQP) (Evert and Hardie 2011) allows complex query patterns and, together with CWB, is typically used as the back-end for web-based corpus interfaces, such as the British National Corpus. Since Version 3.0, CWB offers advanced XML support to import corpora.[2]

[2] http://cwb.sourceforge.net/files/CWB_Encoding_Tutorial/node5.html

Another impactful project is **NAF**, the NLP Annotation Format, a stand-off, multilayered annotation schema for representing linguistic annotations, designed to work well with complex NLP pipelines.[3]

Built on top of the Linguistic Annotation Framework (LAF) (Ide and Romary 2003) and Kyoto Annotation Format (KAF) (Bosma et al. 2009), NAF exploits LAF-based layers to make it easier its usage within NLP architectures. NAF uses URIs extensively and can be converted to RDF-NAF, an alternative representation that can be read by RDF parsers. This provides interoperability and also helps avoiding redundancy when listing the core elements of a XML tree. Furthermore, it enhances the usage of Linked Data by the NLP community.

The NTUMC DTD

The XML template hereby described is compliant to the official **NTUMC DTD**, developed and maintained by the NTU Computational Linguistics Lab.[4]

The purpose of a Document Type Definition (DTD) is to specify the rules that all well-formed XML files need to respect in order to be valid. Therefore, aspects such as uniqueness of the ids have to be guaranteed beforehand, *while compiling the XML*, thus making room for an easier and less error-prone sharing practice in the long run.

Corpora are often monolingual, but the template should be as inclusive and

[3] https://github.com/newsreader/NAF

[4] http://compling.hss.ntu.edu.sg/

general-purpose as possible. Hence, the chosen format should provide the means to include multilingual documents, each with their own language, even allowing more than one document per language to be linked. NTUMC, for instance, is already composed by corpora having different translations in the same language of the same text.

As it follows, the template should be designed to include at least:

- document marking;
- document language;
- sentence marking;
- sentence raw text;
- token (surface form), lemma, PoS (using UPOS) information at the token level;
- if available, sense annotation following WN convention;
- if the corpus is multilingual, alignments at all structural levels.

The goal is to be able to produce any corpus in a single file regardless of how many languages it contains. The motivation for such a format is that managing parallel corpora separately may allow bad practices, like sentence and document ids that fail to match. Nonetheless, the XML format should allow even a multilingual corpus to be released in as many monolingual self-contained XML files as many languages it consists of.

Another major benefit is that anyone will be able to export their corpus to an XML complying to this DTD. This will guarantee that the document is well formed and directly importable into the NTUMC from the start.

Appendix C shows the DTD proposed, which is also accessible at https://github.com/lmorgadodacosta/NTUMC, where it is currently maintained and continuously improved. For the reader's convenience, here follow smaller snippets of the DTD, along with a few words of explanation.

```
<!ELEMENT Corpus (Document+, Alignment?)>
<!ATTLIST Corpus
          corpusID ID #REQUIRED
          title CDATA #REQUIRED
          linguality (multilingual | monolingual) #REQUIRED>
```

Listing 5.1: NTUMC DTD: Root element `Corpus`.

The `Corpus` element is the root of the XML and has to contain at least one `Document` and, optionally, one or more `Alignment` elements; see Listing 5.1.

Language information is expressed at the document level. At the corpus level, attribute `linguality` only states whether the collection of documents is monolingual or multilingual.

Taking advantage of the embedded hierarchical structure of the XML, information such as language is not reiterated for every element, as it can be inferred. The relationships between parent and children nodes are also inherently expressed by the tree structure.

It is mandatory that a unique identifier is assigned to every element. Ids are incrementally produced and type-identified by the first letter. All ″to″ idref attributes are actually idrefs, so all the multiple references are contained in one single element, space-separated.

```
<!ELEMENT Document (Sentence+)>
<!ATTLIST Document
          docID ID #REQUIRED
          doc CDATA #REQUIRED
          language CDATA #REQUIRED
          title CDATA #REQUIRED
          subtitle CDATA #IMPLIED
          url CDATA #IMPLIED
          collection CDATA #IMPLIED>
<!ELEMENT Sentence (Word*, Concept*, Chunk*, Tag*)>
<!ATTLIST Sentence
          sid ID #REQUIRED
          sent CDATA #REQUIRED
          pid CDATA #IMPLIED
          comment CDATA #IMPLIED
          last_changed_by CDATA #IMPLIED
          last_changed CDATA #IMPLIED>
```

Listing 5.2: NTUMC DTD: elements Document and Sentence.

Similarly to the relationship between a corpus and its documents, every Document has to contain at least one Sentence and every Sentence at least one Word (Listing 5.2).

```
<!ELEMENT Word (Tag*)>
<!ATTLIST Word
          wid ID #REQUIRED
          lemma CDATA #IMPLIED
          surface_form CDATA #REQUIRED
          upos (ADJ|ADP|ADV|AUX|CCONJ|DET|INTJ|NOUN|NUM|PART|PRON|
          PROPN|PUNCT|SCONJ|SYM|VERB|X) #IMPLIED
          pos CDATA #IMPLIED
          cfrom CDATA #IMPLIED
          cto CDATA #IMPLIED
          comment CDATA #IMPLIED
          last_changed_by CDATA #IMPLIED
          last_changed CDATA #IMPLIED>
<!ELEMENT Concept (Tag*)>
<!ATTLIST Concept
          cid ID #REQUIRED
          wid IDREFS #REQUIRED
          clemma CDATA #REQUIRED
          synset_tag CDATA #IMPLIED
          comment CDATA #IMPLIED
          last_changed_by CDATA #IMPLIED
          last_changed CDATA #IMPLIED>
<!ELEMENT Chunk (Tag*)>
<!ATTLIST Chunk
          chid ID #REQUIRED
          wid IDREFS #REQUIRED
          comment CDATA #IMPLIED
          last_changed_by CDATA #IMPLIED
          last_changed CDATA #IMPLIED
<!ELEMENT Tag EMPTY>
<!ATTLIST Tag
              category CDATA #REQUIRED
              value CDATA #REQUIRED
              comment CDATA #IMPLIED
              last_changed_by CDATA #IMPLIED
          last_changed CDATA #IMPLIED
          confidence CDATA #IMPLIED>
```

Listing 5.3: NTUMC DTD: elements `Word`, `Concept`, `Chunk` and `Tag`.

`Concept` elements are optional, as it cannot be assumed, a priori, that the corpus to be imported will also have semantic annotation. The same goes for attributes `lemma` and `pos` for element `Word`, as in Listing 5.3.

In order to enhance the reusability of new corpora as much as possible, it would be ideal to make room for universal parts of speech, the standard here being the Universal Part of Speech Tagger (UPOS) (Petrov et al. 2012). For existing corpora already tagged with a language-specific tagset, however, conversion could be taken care of by the pipeline: mappings of very popular tagsets are already available for English, Chinese, Japanese and Indonesian in the NTUMC website. As an example, Table 5.1 shows the mapping between the widely used TANL tagset for Italian and UPOS. As it can be seen, at times the mapping is not straightforward and can lead to loss of information, which is prevented by keeping both tags in the XML.

TANL	UPOS	TANL	UPOS	TANL	UPOS
A	ADJ	E	ADP	R	DET
AP	PRON	F	PUNCT	S	NOUN
B	ADV	I	INTJ	T	ADJ
C	CCONJ | SCONJ	N	NUM	V	VERB | AUX
D	DET	P	PRON	X	X

Table 5.1: Mapping of Italian PoS-tagset TANL onto UPOS

Moving on with the DTD, element Tag allows for another layer within the sentence to be used in a more general way, its purpose being expressed in attribute category. Tag elements can also be children to Word or Concept; a Tag element child to a Concept element is the ideal place to store the annotation obtained through MSI, making also use of attribute confidence to signal that the annotation is not hand-checked.

```
<!ELEMENT Alignment (DocAlignment*, SentAlignment*, WordAlignment*,
    ConceptAlignment*)>
<!ELEMENT DocAlignment EMPTY>
<!ATTLIST DocAlignment
          docID IDREFS #REQUIRED
          type (manual|automatic) #REQUIRED>
<!ELEMENT SentAlignment EMPTY>
<!ATTLIST SentAlignment
          sid_from IDREF #REQUIRED
          sid_to IDREFS #REQUIRED
          type (manual|automatic) #REQUIRED>
<!ELEMENT WordAlignment EMPTY>
<!ATTLIST WordAlignment
          wid_from IDREF #REQUIRED
          wid_to IDREFS #REQUIRED
          type (manual|automatic) #REQUIRED>
<!ELEMENT ConceptAlignment EMPTY>
<!ATTLIST ConceptAlignment
          cid_from IDREF #REQUIRED
          cid_to IDREFS #REQUIRED
          type (manual|automatic) #REQUIRED>
```

Listing 5.4: NTUMC DTD: Alignment element and allowed children nodes.

The fact that `Alignment` elements are children to `Corpus` allows for less reduplication, as the pairs (or tuples) can be defined only once, instead of repeating the information at each `Token` level.

All `Alignment` elements are defined so that the unit of interest allows also for partial matching; see Listing 5.4. For instance, it could be the case that in a multilingual corpus a given English sentence `s1` is aligned to its Italian equivalent, which however happens to be spread over two sentences, `s2` and `s3`. This issue can be overcome by always defining the span of the alignment with the attributes `sid_from` and `sid_to`, with the latter being of type `idrefs` so to allow multiple alignments.

In the future, `Concept` tags will likely include a `ili_tag` attribute along with `synset_tag`, to enhance compatibility with the InterLingual Index (Bond et al. 2016; Peters et al. 1998).

The `chunk` element serves the goal of representing a kind of word clustering different than concepts. Its primary use in the NTUMC corpus is for sentiment tagging and error tagging (the latter for Learner corpora).

`Tag` can be optionally found within sentences, chunks, words and concepts to provide further layers of annotations to the parent element, being used for anything from sentiment, multiword expression tagging, grammar errors, automatic annotations, etc.

The DTD hereby proposed is very barebones, but it can be easily extended to make room for further annotation layers, such as sentiment, grammatical and style tags, verb argument structure, thematic roles thanks to the `Tag` element.

5.1.3 Making MPC available

A fundamental step is to upload the components of the MPC on a public repository on Github, for everyone to benefit from a multilingual parallel corpus with sense annotations aligned to WordNet 3.0. The resources, along with the code used to produce them, are available at `https://github.com/jusing-es/clwsd`.

Appendix D shows a minimal working example of the corpus, i.e. one aligned document in the four languages (consisting of one sentence only due to space constraints), showing the alignments at the document, sentence, word and concept level.

5.2 A pipeline for the NTUMC toolkit: from raw parallel text to sense annotation

Making the DTD available for the community, along with well-formed XML exported from the NTUMC database, allows everyone who desires to contribute to

NTUMC to do so autonomously.

As explained before, this work is to be part of the NTUMC toolkit (Tan and Bond 2014): the current tools should be integrated to accept any parallel corpus in the format just described, perform MSI upon it and then import the output multilingual corpus, enriched with new sense annotations, in the database.

Future work on NTUMC toolkit should equip the current pipeline with tools that take any plain text and progressively analyze and enrich it with all annotation levels. Tools for lemmatization and POS tagging are already available for English, Mandarin Chinese and Japanese, and more for the other NTUMC languages can be integrated (i.e. Indonesian, Italian, etc.).

Word Alignments for NTUMC

Ongoing work is being done to provide the pipeline with a procedure that automatically produce word alignments, if not already available, and store them along in the NTUMC database during the importing step.

The tool takes as input plain unaligned text for each language pair of the parallel corpus and, assuming that the sentence segmentation matches for both sides, converts it into plain aligned text; finally, it exploits `Fast_align`[5] to produce word alignments.

If available for the languages involved, additional training data is used to en-

[5] https://github.com/clab/fast_align

hance the accuracy of the resulting alignment. Training data come from any parallel corpora shared in the academic community with appropriate license and a format that can be made compatible with NTUMC toolkit.

Chapter 6

Conclusions and future work

To our knowledge, this is the first attempt to disambiguate a parallel corpus by using multilingual MSI. The more languages are considered, the more ambiguity should be reduced and the better MSI is expected to perform.

In future work, we plan to perform MSI on a different parallel corpus. NTU Multilingual Corpus (NTUNMC) (Tan and Bond 2014) would make a sensible choice: suitable parts could be either the YourSing section, which consists of tourism-related texts in seven languages, or, on a much smaller scale, *The Adventure of the Speckled Band* by Sir Arthur Conan Doyle, which is also available in Italian.[1] It would be useful to calculate SFS from untagged text, following McCarthy and Carroll (2003).

Furthermore, we are investigating alternative ways to solve the ambiguity left

[1] Spanish, German, Dutch and Polish translations are also being prepared.

whenever MSI does not lead to a single synset; for instance, we plan to apply some implementation of Lesk (Lesk 1986; Petrolito 2016) on the subset found by MSI.

For future work, it is important to dig deeper into understanding the progressive improvement that can be achieved by taking into account semantic information from one language at the time, so as to verify if it is true that it is the very diverse languages that contribute the most to the disambiguation process.

As for the sense inventories, it would be interesting to compare different lexical resources for Italian, that is MWN and ItalWordNet (ITW) (Roventini et al. 2000). ITW was born as the EuroWordNet Italian database, but even though compatible to a certain extent with EuroWordNet, it is released in XML format. ITW includes about 47.000 lemmas, 50.000 synsets and 130.000 semantic relations and is currently maintained by the Institute for Computational Linguistics (ILC) at the National Research Council (CNR). An updated version is freely available online. [2]

The approach hereby described can be applied to any parallel corpus, possibly extending to all the the languages we have wordnets for. Producing new semantically annotated resources would open the gates to many further tasks, such as estimating basic vocabulary, learning MWE patterns, estimating the predominant sense, and so on.

As for future work, as Rada Mihalcea's English Semcor 3.0 is already available as a corpus in NLTK (see http://www.nltk.org/nltk_data/), it would be useful to write a specific corpus reader for every other SemCor sibling, so one could nav-

[2] http://datahub.io/dataset/iwn

igate the original XML trees in more convenient format. This would be especially valuable, as NLTK is lacking resources in languages different than English.

All data and scripts derived by our work have been made available, except for those derived from RSC, as its license currently forbids it.

Appendices

A List of SemCor texts available in all four languages

br-a01
br-a11
br-a12
br-a13
br-a14
br-b13
br-b20
br-c01
br-c02
br-c04
br-d01
br-d02
br-d03
br-e01
br-e04
br-e24
br-e29
br-f03
br-f10
br-f19
br-f43
br-g11
br-g15
br-h01
br-j01
br-j03
br-j04
br-j05
br-j10
br-j17
br-j22
br-j23
br-j37
br-j52
br-j53
br-j55
br-j57
br-j58
br-j60
br-k01
br-k02
br-k03
br-k05
br-k08
br-k10
br-k11
br-k13
br-k15
br-k18
br-k19

B Lemmas found missing in WordNet 3.0

In the following subsections, due to space constraint we report only lemmas found missing at least more than once in WordNet 3.0. The complete lists can be found at https://github.com/jusing-es/clwsd/tree/master/resources/missing_senses.

B.1 English

Lemma	POS	Occurrences	Suggested sense
a_thousand	a	3	02198752-a
a_hundred	a	2	02196107-a

B.2 Italian

Lemma	POS	Occurrences	Translation	Suggested sense
densità	n	7	density	04941453-n
più	r	6	more	00099341-r
termale	a	5	thermal	02814453-a
ingiusto	a	3	unfair	00957176-a
così	r	3	so	00146594-r
più	r	3	most	00111609-r
stazione_di_servizio	n	3	garage	03416489-n
record_mondiale	n	2	world_record	00063559-n
volgare	a	2	vulgar	01950198-a
voluttuoso	a	2	voluptuous	02132967-a
visuale	a	2	visual	02869563-a
vibrante	a	2	vibrant	02280969-a

Utah	n	2	utah	09147046-n
impreparato	a	2	unprepared	01845160-a
impassibile	a	2	unmoved	01560320-a
unanime	a	2	unanimous	00553732-a
volgere	v	2	turn	01907258-v
triplo	n	2	triple	00132982-n
Texas	n	2	texas	09141526-n
sottendere	v	2	subtend	02693842-v
risvegliare	v	2	stir	01761706-v
aderire	v	2	stick_to	01356750-v
organico	n	2	staff	08439955-n
struttura_sociale	n	2	social_structure	08378819-n
liscio	a	2	smooth	02236842-a
esiguo	a	2	small	01391351-a
sedicesimo	a	2	sixteenth	02204131-a
laterale	a	2	side	08649345-n
sic	r	2	sic	00146500-r
settimo	a	2	seventh	02202979-a
serratus	n	2	serratus	05550330-n
righthander	n	2	right-hander	10387324-n
reversibile	a	2	reversible	01758934-a
restituire	v	2	repay	02284951-v
reliquia	n	2	relic	04073547-n
registrato	a	2	recorded	01422956-a
rassicurare	v	2	reassure	01766407-v
giungere	v	2	reach	00743344-v
razionalizzare	v	2	rationalize	00894738-v
viola	a	2	purple	00380312-a
pianificatore	n	2	planner	10438172-n

pittoresco	a	2	picturesque	00219924-a
Philadelphia	n	2	philadelphia	09136182-n
Phase	n	2	phase	15290337-n
par	n	2	par	13596756-n
operistico	a	2	operatic	02912383-a
nazionalistico	a	2	nationalistic	01740358-a
muscoloso	a	2	muscular	00828336-a
Montreal	n	2	montreal	08829533-n
mancante	a	2	missing	02127853-v
Massachusetts	n	2	massachusetts	09095023-n
Corpo_della_Marina	n	2	marine_corps	08192970-n
manoscritto	n	2	manuscript	06406979-n
perso	a	2	lost	01450969-a
vicegovernatore	n	2	lieutenant_governor	10260322-n
colto	a	2	learned	02084358-a
ionizzato	a	2	ionized	00356110-a
integrante	a	2	integral	01348528-a
inibitore	a	2	inhibitory	02004176-a
idrofobo	a	2	hydrophobic	00491749-a
umiliante	a	2	humiliating	00752555-a
Houston	n	2	houston	09144851-n
scuola_media_superiore	n	2	high_school	08409617-n
ibernare	v	2	hibernate	00015946-v
ecco	r	2	here	00108773-r
Harlem	n	2	harlem	09121334-n
ginnastico	a	2	gymnastic	00032497-a
ginnico	a	2	gymnastic	00032497-a
guanto	n	2	glove	02800213-n
Gibson_Girl	n	2	gibson_girl	10129338-n

fullback	n	2	fullback	10115430-n
formarsi	v	2	form	02623906-v
lungi	r	2	far	00101323-r
fairway	n	2	fairway	08569319-n
esteso	a	2	extensive	01386234-a
previsto	a	2	expected	00929567-a
ciascuno	a	2	every	02269794-a
ottanta	a	2	eighty	02194151-a
orientale	a	2	eastern	00823971-a
drammaticamente	r	2	dramatically	00138945-r
distintivo	a	2	distinctive	00357556-a
dissolvere	v	2	dispel	02002720-v
diminuito	a	2	diminished	01274945-a
fioco	a	2	dim	00275290-a
vice	n	2	deputy	10005548-n
dentale	a	2	dental	02711098-a
deltoide	n	2	deltoid	05549350-n
tagliato	a	2	cut	00661278-a
sgualcito	a	2	crushed	02240668-a
grossolano	a	2	crude	02229584-a
cps	n	2	cps	15279104-n
convulsivo	a	2	convulsive	02303754-a
conservativo	a	2	conservative	00574422-a
riflesso_condizionato	n	2	conditioned_reflex	00864226-n
clubhouse	n	2	clubhouse	03054311-n
chiaro	r	2	clearly	00039058-r
rivendicazione	n	2	claim	06729864-n
cinematografico	a	2	cinematic	02696795-a
Padri_della_Chiesa	n	2	church_father	09921792-n

cerebrale	a	2	cerebral	01927455-a
bodybuilders	n	2	bodybuilder	09862845-n
biologico	a	2	biological	02665803-a
piegato	a	2	bent	06199702-n
Bench_Press	n	2	bench_press	00626574-n
barbuto	a	2	bearded	02153965-a
auditorium	n	2	auditorium	02758134-n
presumere	v	2	assume	00632236-v
autorizzato	a	2	approved	00179035-a
analettico	a	2	analeptic	02309800-a
alkali_bee	n	2	alkali_bee	02210921-n
fungere	v	2	act_as	02671613-v

B.3 Romanian

Lemma	POS	Occurrences	Translation	Suggested sense
de_fapt	r	9	actually	00149510-r
în_general	r	8	generally	00155621-r
deja_vu	n	7	deja_vu	05810440-n
în_același_timp	r	7	at_the_same_time	00120095-r
de_asemenea	r	78	also	00047534-r
avea_nevoie	v	6	need	01188725-v
corp_negru	n	6	blackbody	09222406-n
și_așa_mai_departe	r	6	and_so_on	00103664-r
din_păcate	r	5	unfortunately	00042769-r
în_general	r	5	in_general	00041954-r
în_jos	r	5	down	00095320-r
de_asemenea	r	5	as_well	00047534-r

fi_de_acord	v	5	agree	00805376-v
în_sus	r	4	up	00096333-r
fără_îndoială	r	4	undoubtedly	00079107-r
deget_de_la_picior	n	4	toe	05577410-n
punct_de_vedere	n	4	standpoint	06210363-n
vânzare_cu_amănuntul	n	4	retailing	01115866-n
energie_potențială	n	4	potential_energy	11494472-n
da_din_cap	v	4	nod	00898434-v
între_timp	r	4	meanwhile	00065184-r
pune_în_scenă	v	3	stage	01711445-v
radiație_solară	n	3	solar_radiation	11510067-n
pur_și_simplu	r	3	simply	00246296-r
doi	a	3	latter	01047561-a
stat_în_mâini	n	3	handstand	00436187-n
în_final	r	3	finally	00065822-r
în_final	r	3	finally	00047903-r
teme	v	3	fear	01780729-v
duzină	n	3	dozen	13746785-n
război_rece	n	3	cold_war	13982000-n
mic_dejun	n	3	breakfast	07574602-n
sârmă_ghimpată	n	3	barbed_wire	02790823-n
albină_lucrătoare	n	2	worker_bee	02207805-n
în_sus	r	2	upwards	00096333-r
în_sus	r	2	upward	00096333-r
lua_parte	v	2	take_part	02450256-v
balansa	v	2	swing	01877355-v
temperatura_camerei	n	2	room_temperature	05014442-n
de_fapt	r	2	really	00149510-r
viață_reală	n	2	real_life	05810250-n

pictură_în_ulei	n	2	oil_painting	03844349-n
club_de_noapte	n	2	nightclub	02931417-n
cădere_nervoasă	n	2	nervous_breakdown	14066661-n
rezonanță_magnetică	n	2	magnetic_resonance	11478682-n
în_față	r	2	in_front	00066781-r
emisie_infraroșie	n	2	infrared_emission	11469481-n
ființă_umană	n	2	human_being	02472293-n
înaltă_fidelitate	n	2	high_fidelity	01020488-n
jucător_de_golf	n	2	golfer	10136959-n
din_fericire	r	2	fortunately	00042254-r
grădină_de_flori	n	2	flower_garden	03368637-n
la_fel_de	r	2	equally	00022131-r
devota	v	2	devote	00887463-v
definit	a	2	definite	00700451-a
răci	v	2	cool	00370412-v
reflex_condiționat	n	2	conditioned_reflex	00864226-n
cortex_cerebral	n	2	cerebral_cortex	05486510-n
ivi	v	2	arise	02624263-v
arcui	v	2	arch	02034986-v
în_față	r	2	ahead	00066781-r

B.4 Japanese

Lemma	POS	Occurrences	Translation	Suggested sense
使用	v	20	use	01158872-v
提供	v	20	provide	02327200-v
存在	v	17	exist	02603699-v
利用	v	16	use	01158872-v

座る	v	16	sit	01543123-v
説明	v	14	explain	00939277-v
トレーニング	n	12	training	00893955-n
所有	v	8	own	02204692-v
警告	v	7	warn	00870213-v
理解	v	7	understand	00588888-v
除去	v	7	remove	00173338-v
拒否	v	7	refuse	00797430-v
識別	v	7	recognize	02193194-v
提供	v	7	offer	02296726-v
議論	v	7	discuss	01034312-v
チャンピオン	n	7	champion	09906704-n
電話	v	7	call	00789448-v
要求	v	6	require	02627934-v
到達	v	6	reach	02020590-v
練習	v	6	practice	00606093-v
プレー	v	6	play	01072949-v
殺害	v	6	kill	01323958-v
改善	v	6	improve	00205885-v
比較	v	6	compare	00652900-v
影響	v	6	affect	00137313-v
達成	v	6	achieve	02526085-v
心配	v	5	worry	01767163-v
得点	v	5	score	01111816-v
報告	v	5	report	00966809-v
推薦	v	5	recommend	00875141-v
上演	v	5	present	01711445-v
準備	v	5	prepare	00406243-v
ページ	n	5	page	06256697-n

交尾	v	5	mate	01428853-v
着地	v	5	land	01979901-v
発行	v	5	issue	00967625-v
開催	v	5	hold	01733477-v
予期	v	5	expect	00719734-v
否定	v	5	deny	00817003-v
対処	v	5	deal	02587532-v
コミューン	n	5	commune	08541609-n
コーヒー	n	5	coffee	07929519-n
存在	v	5	be	02603699-v
ベースボール	n	5	baseball	00471613-n

C NTUMC DTD

```
<!ELEMENT Corpus (Document+, Alignment?)>
<!ATTLIST Corpus
          corpusID ID #REQUIRED
          title CDATA #REQUIRED
          linguality (multilingual | monolingual) #REQUIRED>
<!ELEMENT Document (Sentence+)>
<!ATTLIST Document
          docID ID #REQUIRED
          doc CDATA #REQUIRED
          language CDATA #REQUIRED
          title CDATA #REQUIRED
          subtitle CDATA #IMPLIED
          url CDATA #IMPLIED
          collection CDATA #IMPLIED>
<!ELEMENT Tag EMPTY>
<!ATTLIST Tag
              category CDATA #REQUIRED
              value CDATA #REQUIRED
              comment CDATA #IMPLIED
              last_changed_by CDATA #IMPLIED
          last_changed CDATA #IMPLIED
          confidence CDATA #IMPLIED>
<!ELEMENT Sentence (Word*, Concept*, Chunk*, Tag*)>
<!ATTLIST Sentence
          sid ID #REQUIRED
          sent CDATA #REQUIRED
          pid CDATA #IMPLIED
          comment CDATA #IMPLIED
          last_changed_by CDATA #IMPLIED
          last_changed CDATA #IMPLIED>
<!ELEMENT Word (Tag*)>
<!-- Allowed values for UPOS: http://universaldependencies.org/u/pos/
   Universal Pos Tag -->
<!ATTLIST Word
          wid ID #REQUIRED
          lemma CDATA #IMPLIED
          surface_form CDATA #REQUIRED
          upos (ADJ|ADP|ADV|AUX|CCONJ|DET|INTJ|NOUN|NUM|PART|PRON|
          PROPN|PUNCT|SCONJ|SYM|VERB|X) #IMPLIED
          pos CDATA #IMPLIED
          cfrom CDATA #IMPLIED
          cto CDATA #IMPLIED
          comment CDATA #IMPLIED
          last_changed_by CDATA #IMPLIED
          last_changed CDATA #IMPLIED>
<!ELEMENT Concept (Tag*)>
<!ATTLIST Concept
          cid ID #REQUIRED
          wid IDREFS #REQUIRED
          clemma CDATA #REQUIRED
          synset_tag CDATA #IMPLIED
          comment CDATA #IMPLIED
          last_changed_by CDATA #IMPLIED
          last_changed CDATA #IMPLIED>
<!ELEMENT Chunk (Tag*)>
<!ATTLIST Chunk
          chid ID #REQUIRED
          wid IDREFS #REQUIRED
          comment CDATA #IMPLIED
          last_changed_by CDATA #IMPLIED
          last_changed CDATA #IMPLIED>
<!ELEMENT Alignment (DocAlignment*, SentAlignment*, WordAlignment*,
   ConceptAlignment*)>
<!ELEMENT DocAlignment EMPTY>
<!ATTLIST DocAlignment
          docID IDREFS #REQUIRED
          type (manual|automatic) #REQUIRED>
<!ELEMENT SentAlignment EMPTY>
<!ATTLIST SentAlignment
          sid_from IDREF #REQUIRED
          sid_to IDREFS #REQUIRED
          type (manual|automatic) #REQUIRED>
<!ELEMENT WordAlignment EMPTY>
<!ATTLIST WordAlignment
          wid_from IDREF #REQUIRED
          wid_to IDREFS #REQUIRED
          type (manual|automatic) #REQUIRED>
<!ELEMENT ConceptAlignment EMPTY>
<!ATTLIST ConceptAlignment
          cid_from IDREF #REQUIRED
          cid_to IDREFS #REQUIRED
          type (manual|automatic) #REQUIRED>
```

D Sample from the Multilingual Parallel Corpus

```
<?xml version=”1.0” encoding=”UTF−8”?>
<!DOCTYPE Corpus SYSTEM ”../../NTUMC/ntumc.dtd”>
<Corpus corpusID=”MPC” title=”Multilingual Parallel Corpus” linguality=”multilingual”>
        <Document docID=”eng_a01” doc=”text” language=”eng” title=”a01”>
            <Sentence sid=”eng_s_1” sent=”Fulton_County_Grand_Jury said Friday investigation Atlanta recent
                  primary_election produced evidence irregularities took_place”>
                <Word wid=”eng_t_1_2” pos=”n” lemma=”group” surface_form=”Fulton_County_Grand_Jury”/>
                <Word wid=”eng_t_1_3” pos=”v” lemma=”say” surface_form=”said”/>
                <Word wid=”eng_t_1_4” pos=”n” lemma=”friday” surface_form=”Friday”/>
                <Word wid=”eng_t_1_6” pos=”n” lemma=”investigation” surface_form=”investigation”/>
                <Word wid=”eng_t_1_8” pos=”n” lemma=”atlanta” surface_form=”Atlanta”/>
                <Word wid=”eng_t_1_10” pos=”a” lemma=”recent” surface_form=”recent”/>
                <Word wid=”eng_t_1_11” pos=”n” lemma=”primary_election” surface_form=”primary_election”/>
                <Word wid=”eng_t_1_12” pos=”v” lemma=”produce” surface_form=”produced”/>
                <Word wid=”eng_t_1_15” pos=”n” lemma=”evidence” surface_form=”evidence”/>
                <Word wid=”eng_t_1_19” pos=”n” lemma=”irregularity” surface_form=”irregularities”/>
                <Word wid=”eng_t_1_20” pos=”v” lemma=”take_place” surface_form=”took_place”/>
                <Concept cid=”eng_c_1_2” wid=”eng_t_1_2” synset_tag=”00031264−n” clemma=”group”/>
                 <Concept cid=”eng_c_1_3” wid=”eng_t_1_3” synset_tag=”01009240−v” clemma=”say”/>
                 <Concept cid=”eng_msi_c_1_3” wid=”eng_t_1_3” synset_tag=”01016002−v” clemma=”say”>
                    <Tag category=”msi_annotation” value=”01016002−v” />
                </Concept>
                 <Concept cid=”eng_c_1_4” wid=”eng_t_1_4” synset_tag=”15164463−n” clemma=”friday”/>
                 <Concept cid=”eng_msi_c_1_4” wid=”eng_t_1_4” synset_tag=”15164463−n” clemma=”friday”>
                    <Tag category=”msi_annotation” value=”15164463−n” />
                </Concept>
                 <Concept cid=”eng_c_1_6” wid=”eng_t_1_6” synset_tag=”05800611−n” clemma=”investigation”/>
                 <Concept cid=”eng_msi_c_1_6” wid=”eng_t_1_6” synset_tag=”00633864−n” clemma=”investigation”>
                    <Tag category=”msi_annotation” value=”00633864−n” />
                </Concept>
                 <Concept cid=”eng_c_1_8” wid=”eng_t_1_8” synset_tag=”09076675−n” clemma=”atlanta”/>
                 <Concept cid=”eng_msi_c_1_8” wid=”eng_t_1_8” synset_tag=”09076675−n” clemma=”atlanta”>
                    <Tag category=”msi_annotation” value=”09076675−n” />
                </Concept>
                 <Concept cid=”eng_c_1_10” wid=”eng_t_1_10” synset_tag=”01730444−a” clemma=”recent”/>
                 <Concept cid=”eng_msi_c_1_10” wid=”eng_t_1_10” synset_tag=”01642477−a” clemma=”recent”>
                    <Tag category=”msi_annotation” value=”01642477−a” />
                </Concept>
                 <Concept cid=”eng_c_1_11” wid=”eng_t_1_11” synset_tag=”00182571−n” clemma=”primary_election”/>
                 <Concept cid=”eng_msi_c_1_11” wid=”eng_t_1_11” synset_tag=”00182571−n” clemma=”primary_election”>
                    <Tag category=”msi_annotation” value=”00182571−n” />
                </Concept>
                 <Concept cid=”eng_c_1_12” wid=”eng_t_1_12” synset_tag=”02141146−v” clemma=”produce”/>
                 <Concept cid=”eng_msi_c_1_12” wid=”eng_t_1_12” synset_tag=”01752884−v” clemma=”produce”>
                    <Tag category=”msi_annotation” value=”01752884−v” />
                </Concept>
```

```
           <Concept cid=”eng_c_1_15” wid=”eng_t_1_15” synset_tag=”05823932−n” clemma=”evidence”/>
           <Concept cid=”eng_msi_c_1_15” wid=”eng_t_1_15” synset_tag=”05823932−n” clemma=”evidence”>
              <Tag category=”msi_annotation” value=”05823932−n” />
          </Concept>
           <Concept cid=”eng_c_1_19” wid=”eng_t_1_19” synset_tag=”00737188−n” clemma=”irregularity”/>
           <Concept cid=”eng_msi_c_1_19” wid=”eng_t_1_19” synset_tag=”04770211−n” clemma=”irregularity”>
              <Tag category=”msi_annotation” value=”04770211−n” />
          </Concept>
           <Concept cid=”eng_c_1_20” wid=”eng_t_1_20” synset_tag=”00339934−v” clemma=”take_place”/>
           <Concept cid=”eng_msi_c_1_20” wid=”eng_t_1_20” synset_tag=”00339934−v” clemma=”take_place”>
              <Tag category=”msi_annotation” value=”00339934−v” />
          </Concept>
           </Sentence>
</Document>
  <Document docID=”ita_a01” doc=”text” language=”ita” title=”a01”>
      <Sentence sid=”ita_s_1” sent=”Venerdì detto indagine recente Atlanta prodotto prova verificate irregolarità
           ”>
          <Word wid=”ita_t_1_1” pos=”n” lemma=”venerdì” surface_form=”Venerdì”/>
          <Word wid=”ita_t_1_5” pos=”v” lemma=”dire” surface_form=”detto”/>
          <Word wid=”ita_t_1_8” pos=”n” lemma=”indagine” surface_form=”indagine”/>
          <Word wid=”ita_t_1_10” pos=”a” lemma=”recente” surface_form=”recente”/>
          <Word wid=”ita_t_1_14” pos=”n” lemma=”Atlanta” surface_form=”Atlanta”/>
          <Word wid=”ita_t_1_17” pos=”v” lemma=”produrre” surface_form=”prodotto”/>
          <Word wid=”ita_t_1_20” pos=”n” lemma=”prova” surface_form=”prova”/>
          <Word wid=”ita_t_1_27” pos=”v” lemma=”verificare” surface_form=”verificate”/>
          <Word wid=”ita_t_1_29” pos=”n” lemma=”irregolarità” surface_form=”irregolarità”/>
          <Concept cid=”ita_c_1_1” wid=”ita_t_1_1” synset_tag=”15164463−n” clemma=”venerdì”/>
           <Concept cid=”ita_msi_c_1_1” wid=”ita_t_1_1” synset_tag=”15164463−n” clemma=”venerdì”>
              <Tag category=”msi_annotation” value=”15164463−n” />
          </Concept>
           <Concept cid=”ita_c_1_5” wid=”ita_t_1_5” synset_tag=”01009240−v” clemma=”dire”/>
           <Concept cid=”ita_msi_c_1_5” wid=”ita_t_1_5” synset_tag=”01009240−v” clemma=”dire”>
              <Tag category=”msi_annotation” value=”01009240−v” />
          </Concept>
           <Concept cid=”ita_c_1_8” wid=”ita_t_1_8” synset_tag=”05800611−n” clemma=”indagine”/>
           <Concept cid=”ita_msi_c_1_8” wid=”ita_t_1_8” synset_tag=”05800611−n” clemma=”indagine”>
              <Tag category=”msi_annotation” value=”05800611−n” />
          </Concept>
           <Concept cid=”ita_c_1_10” wid=”ita_t_1_10” synset_tag=”01730444−a” clemma=”recente”/>
           <Concept cid=”ita_msi_c_1_10” wid=”ita_t_1_10” synset_tag=”01642477−a” clemma=”recente”>
              <Tag category=”msi_annotation” value=”01642477−a” />
          </Concept>
           <Concept cid=”ita_c_1_14” wid=”ita_t_1_14” synset_tag=”09076675−n” clemma=”Atlanta”/>
           <Concept cid=”ita_msi_c_1_14” wid=”ita_t_1_14” synset_tag=”09076675−n” clemma=”Atlanta”>
              <Tag category=”msi_annotation” value=”09076675−n” />
          </Concept>
           <Concept cid=”ita_c_1_17” wid=”ita_t_1_17” synset_tag=”02141146−v” clemma=”produrre”/>
           <Concept cid=”ita_msi_c_1_17” wid=”ita_t_1_17” synset_tag=”01752884−v” clemma=”produrre”>
              <Tag category=”msi_annotation” value=”01752884−v” />
```

```
            </Concept>
             <Concept cid=”ita_c_1_20” wid=”ita_t_1_20” synset_tag=”05823932−n” clemma=”prova”/>
             <Concept cid=”ita_msi_c_1_20” wid=”ita_t_1_20” synset_tag=”05823932−n” clemma=”prova”>
                <Tag category=”msi_annotation” value=”05823932−n” />
            </Concept>
             <Concept cid=”ita_c_1_27” wid=”ita_t_1_27” synset_tag=”00339934−v” clemma=”verificare”/>
             <Concept cid=”ita_msi_c_1_27” wid=”ita_t_1_27” synset_tag=”02520997−v” clemma=”verificare”>
                <Tag category=”msi_annotation” value=”02520997−v” />
            </Concept>
             <Concept cid=”ita_c_1_29” wid=”ita_t_1_29” synset_tag=”00737188−n” clemma=”irregolarità”/>
             <Concept cid=”ita_msi_c_1_29” wid=”ita_t_1_29” synset_tag=”04770211−n” clemma=”irregolarità”>
                <Tag category=”msi_annotation” value=”04770211−n” />
            </Concept>
             </Sentence>
  </Document>
    <Document docID=”ron_a01” doc=”text” language=”ron” title=”a01”>
        <Sentence sid=”ron_s_1” sent=”Fulton spus vineri alegerilor produs dovadă”>
            <Word wid=”ron_t_1_4” pos=”n” lemma=”group” surface_form=”Fulton”/>
            <Word wid=”ron_t_1_6” pos=”v” lemma=”spune” surface_form=”spus”/>
            <Word wid=”ron_t_1_7” pos=”n” lemma=”vineri” surface_form=”vineri”/>
            <Word wid=”ron_t_1_12” pos=”n” lemma=”alegere” surface_form=”alegerilor”/>
            <Word wid=”ron_t_1_16” pos=”v” lemma=”produce” surface_form=”produs”/>
            <Word wid=”ron_t_1_19” pos=”n” lemma=”dovadă” surface_form=”dovadă”/>
            <Concept cid=”ron_c_1_4” wid=”ron_t_1_4” synset_tag=”00031264−n” clemma=”group”/>
             <Concept cid=”ron_c_1_6” wid=”ron_t_1_6” synset_tag=”01009240−v” clemma=”spune”/>
             <Concept cid=”ron_msi_c_1_6” wid=”ron_t_1_6” synset_tag=”00683771−v” clemma=”spune”>
                <Tag category=”msi_annotation” value=”00683771−v” />
            </Concept>
             <Concept cid=”ron_c_1_7” wid=”ron_t_1_7” synset_tag=”15164463−n” clemma=”vineri”/>
             <Concept cid=”ron_msi_c_1_7” wid=”ron_t_1_7” synset_tag=”15164463−n” clemma=”vineri”>
                <Tag category=”msi_annotation” value=”15164463−n” />
            </Concept>
             <Concept cid=”ron_c_1_12” wid=”ron_t_1_12” synset_tag=”00181781−n” clemma=”alegere”/>
             <Concept cid=”ron_c_1_16” wid=”ron_t_1_16” synset_tag=”02141146−v” clemma=”produce”/>
             <Concept cid=”ron_msi_c_1_16” wid=”ron_t_1_16” synset_tag=”02141146−v” clemma=”produce”>
                <Tag category=”msi_annotation” value=”02141146−v” />
            </Concept>
             <Concept cid=”ron_c_1_19” wid=”ron_t_1_19” synset_tag=”05824739−n” clemma=”dovadă”/>
             </Sentence>
  </Document>
<Document docID=”jpn_a01” doc=”text” language=”jpn” title=”a01”>
        <Sentence sid=”jpn_s_1” sent=”金曜日 いる 示す アトランタ 調査 徴候 予備 選挙 最近 いる 陳べる いる 行なう
              れる 不正 行為”>
            <Word wid=”jpn_w1.1.6” pos=”n” lemma=”金曜日” surface_form=”金曜日”/>
            <Word wid=”jpn_w1.1.44” pos=”v” lemma=”いる” surface_form=”いる”/>
            <Word wid=”jpn_w1.1.36” pos=”v” lemma=”示す” surface_form=”示す”/>
            <Word wid=”jpn_w1.1.8” pos=”n” lemma=”アトランタ” surface_form=”アトランタ”/>
            <Word wid=”jpn_w1.1.15” pos=”n” lemma=”調査” surface_form=”調査”/>
            <Word wid=”jpn_w1.1.32” pos=”n” lemma=”徴候” surface_form=”徴候”/>
            <Word wid="jpn_w1.1.12" pos="n" lemma="予備選挙" surface_form="予備 選挙"/>
            <Word wid="jpn_w1.1.10" pos="n" lemma="最近" surface_form="最近"/>
            <Word wid="jpn_w1.1.38" pos="v" lemma="いる" surface_form="いる"/>
            <Word wid="jpn_w1.1.42" pos="v" lemma="陳べる" surface_form="陳べる"/>
            <Word wid="jpn_w1.1.25" pos="v" lemma="いる" surface_form="いる"/>
            <Word wid="jpn_w1.1.22" pos="v" lemma="行なわれる" surface_form="行なう れる"/>
            <Word wid="jpn_w1.1.19" pos="n" lemma="不正行為" surface_form="不正 行為"/>
            <Concept cid="jpn_c1.1.6" wid="jpn_w1.1.6" synset_tag="15164463−n" clemma="金曜日"/>
             <Concept cid="jpn_msi_w1.1.6" wid="jpn_w1.1.6" synset_tag="15164463−n" clemma="金曜日">
                <Tag category="msi_annotation" value="15164463−n" />
            </Concept>
             <Concept cid="jpn_c1.1.8" wid="jpn_w1.1.8" synset_tag="09076675−n" clemma="アトランタ"/>
             <Concept cid="jpn_msi_w1.1.8" wid="jpn_w1.1.8" synset_tag="09076675−n" clemma="アトランタ">
                <Tag category="msi_annotation" value="09076675−n" />
            </Concept>
             <Concept cid="jpn_c1.1.15" wid="jpn_w1.1.15" synset_tag="05800611−n" clemma="調査"/>
             <Concept cid="jpn_msi_w1.1.15" wid="jpn_w1.1.15" synset_tag="00141806−n" clemma="調査">
                <Tag category="msi_annotation" value="00141806−n" />
            </Concept>
             <Concept cid="jpn_c1.1.12" wid="jpn_w1.1.12" synset_tag="00182571−n" clemma="予備選挙"/>
             <Concept cid="jpn_msi_w1.1.12" wid="jpn_w1.1.12" synset_tag="00182571−n" clemma="予備選挙">
                <Tag category="msi_annotation" value="00182571−n" />
            </Concept>
             <Concept cid="jpn_c1.1.42" wid="jpn_w1.1.42" synset_tag="01009240−v" clemma="陳べる"/>
             <Concept cid="jpn_c1.1.22" wid="jpn_w1.1.22" synset_tag="00339934−v" clemma="行なわれる"/>
             <Concept cid="jpn_msi_w1.1.22" wid="jpn_w1.1.22" synset_tag="00339934−v" clemma="行なわれる">
                <Tag category="msi_annotation" value="00339934−v" />
            </Concept>
             <Concept cid="jpn_c1.1.19" wid="jpn_w1.1.19" synset_tag="00745637−n" clemma="不正行為"/>
             </Sentence>
</Document>
    <Alignment>
    <DocAlignment docID="ita_a01 eng_a01 jpn_a01 ron_a01"/>
        <SentAlignment sid_from="eng_s_1" sid_to="ita_s_1 ron_s_1 jpn_s_1" type="manual"/>
        <SentAlignment sid_from="ita_s_1" sid_to="eng_s_1 ron_s_1 jpn_s_1" type="manual"/>
        <SentAlignment sid_from="ron_s_1" sid_to="eng_s_1 ita_s_1 jpn_s_1" type="manual"/>
        <SentAlignment sid_from="jpn_s_1" sid_to="eng_s_1 ita_s_1 ron_s_1" type="manual"/>
        <WordAlignment wid_from="eng_t_1_3" wid_to="ita_t_1_5 ron_t_1_6 jpn_w1.1.42" type="manual"/>
        <WordAlignment wid_from="eng_t_1_4" wid_to="ita_t_1_1 jpn_w1.1.6" type="manual"/>
        <WordAlignment wid_from="eng_t_1_6" wid_to="ita_t_1_8 jpn_w1.1.15" type="manual"/>
        <WordAlignment wid_from="eng_t_1_8" wid_to="ita_t_1_14 jpn_w1.1.8" type="manual"/>
        <WordAlignment wid_from="eng_t_1_10" wid_to="ita_t_1_10 jpn_w1.1.10" type="manual"/>
        <WordAlignment wid_from="eng_t_1_11" wid_to="jpn_w1.1.12" type="manual"/>
        <WordAlignment wid_from="eng_t_1_12" wid_to="ita_t_1_17 ron_t_1_16" type="manual"/>
        <WordAlignment wid_from="eng_t_1_15" wid_to="ita_t_1_20" type="manual"/>
        <WordAlignment wid_from="eng_t_1_19" wid_to="ita_t_1_29" type="manual"/>
        <WordAlignment wid_from="eng_t_1_20" wid_to="ita_t_1_27 jpn_w1.1.22" type="manual"/>
        <ConceptAlignment cid_from="eng_c_1_8" cid_to="ita_c_1_14 jpn_w1.1.8" type="manual"/>
        <ConceptAlignment cid_from="eng_c_1_3" cid_to="ita_c_1_5 ron_c_1_6 jpn_w1.1.42" type="manual"/>
        <ConceptAlignment cid_from=”eng_c_1_4” cid_to=”ita_c_1_1 jpn_w1.1.6 ron_c_1_7” type=”manual”/>
        <ConceptAlignment cid_from=”eng_c_1_6” cid_to=”ita_c_1_8 jpn_w1.1.15” type=”manual”/>
    </Alignment>
</Corpus>
```

Bibliography

Agirre, Eneko and Philip Edmonds (2007). *Word sense disambiguation: Algorithms and applications*. Vol. 33. Springer Science & Business Media.

Agirre, Eneko and Aitor Soroa (2009). "Personalizing pagerank for word sense disambiguation". In: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 33–41.

Agirre, Eneko and Mark Stevenson (2006). "Knowledge Sources for WSD". In: *Word Sense Disambiguation: Algorithms and Applications*. Ed. by Eneko Agirre and Philip Edmonds. Dordrecht: Springer Netherlands, pp. 217–251.

Agirre, Eneko, Lluís Màrquez, and Richard Wicentowski, eds. (2007). *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Prague, Czech Republic: Association for Computational Linguistics.

Agirre, Eneko et al., eds. (2012). **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Seman-*

tic Evaluation (SemEval 2012). Montréal, Canada: Association for Computational Linguistics.

Bandyopadhyay, S. (2012). *Emerging Applications of Natural Language Processing: Concepts and New Research: Concepts and New Research*. Information Science Reference. ISBN: 9781466621701.

Bentivogli, Luisa and Emanuele Pianta (2002). "Opportunistic Semantic Tagging". In: *Third International Conference on Language Resources and Evaluation (LREC 2002)*, pp. 1401–1406.

— (2005). "Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus". In: *Natural Language Engineering* 11.03, p. 247.

Bentivogli, Luisa and Emanuelle Pianta (2000). "Looking for lexical gaps". In: *Proceedings of the Ninth EURALEX International Congress, EURALEX 2000: Stuttgart, Germany, August 8th-12th, 2000*, pp. 663–669.

Bentivogli, Luisa et al. (2004). "Revising the wordnet domains hierarchy: semantics, coverage and balancing". In: *Proceedings of the Workshop on Multilingual Linguistic Ressources*. Association for Computational Linguistics, pp. 101–108.

Bhattacharya, Indrajit, Lise Getoor, and Yoshua Bengio (2004). "Unsupervised sense disambiguation using bilingual probabilistic models". In: *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, p. 287.

Bird, Steven and Edward Loper (2004). "NLTK: the natural language toolkit". In: *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*. Association for Computational Linguistics, p. 31.

Bonansinga, Giulia and Francis Bond (2016). “Multilingual Sense Intersection in a Parallel Corpus with Diverse Language Families”. In: *Proceedings of the Eighth Global WordNet Conference*. Editura Universităţii Al.I. Cuza din Iaşi (UAIC), pp. 44–49.

Bond, Francis and Giulia Bonansinga (2015). “Exploring Cross-Lingual Sense Mapping in a Multilingual Parallel Corpus”. In: *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015*. Academia University Press, pp. 56–61.

Bond, Francis and Ryan Foster (2013). “Linking and Extending an Open Multilingual Wordnet.” In: *ACL (1)*, pp. 1352–1362.

Bond, Francis and Kyonghee Paik (2012). “A survey of wordnets and their licenses”. In: *Proceedings of the 6th International Global WordNet Conference*, pp. 64–71.

Bond, Francis et al. (2009). “Enhancing the japanese wordnet”. In: *Proceedings of the 7th workshop on Asian language resources*. Association for Computational Linguistics, pp. 1–8.

Bond, Francis et al. (2012). “Japanese SemCor: A sense-tagged corpus of Japanese”. In: *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, pp. 56–63.

Bond, Francis et al. (2013). “Developing Parallel Sense-tagged Corpora with Wordnets”. In: *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 149–158.

Bond, Francis et al. (2016). “CILI: the Collaborative Interlingual Index”. In: *Proc. of the Eighth Global WordNet Conference (GWC 2016)*, pp. 50–57.

Bosma, Wauter et al. (2009). “KAF: a Generic Semantic Annotation Format”. In: pp. 1–8.

Brants, Thorsten (2000). “TnT: a statistical part-of-speech tagger”. In: *Proceedings of the sixth conference on Applied natural language processing*. Association for Computational Linguistics, pp. 224–231.

Breen, James (2004). “JMDict: a Japanese-multilingual dictionary”. In: *Proceedings of the Workshop on Multilingual Linguistic Ressources*. Association for Computational Linguistics, pp. 71–79.

Brill, Eric (1992). “A Simple Rule-based Part of Speech Tagger”. In: *Proceedings of the Third Conference on Applied Natural Language Processing*. ANLC ’92. Trento, Italy: Association for Computational Linguistics, pp. 152–155.

Brown, Peter F. et al. (1991). “Word-Sense Disambiguation Using Statistical Methods”. In: *Proceedings of the 29th Annual Meeting of the ACL*. Morristown, NJ: Morristown, NJ: Association for Computational Linguistics.

Carpuat, Marine and Dekai Wu (2007). “Improving Statistical Machine Translation Using Word Sense Disambiguation.” In: *EMNLP-CoNLL*. Vol. 7, pp. 61–72.

Chan, Yee Seng and Hwee Tou Ng (2005). “Scaling up word sense disambiguation via parallel texts”. In: *AAAI*. Vol. 5, pp. 1037–1042.

Costa, Luís Morgado da and Francis Bond (2015). “OMWEdit-the integrated open multilingual wordnet editing system”. In: *ACL-IJCNLP 2015*, p. 73.

Dagan, Ido, Alon Itai, and Ulrike Schwall (1991). “Two languages are more informative than one”. In: *Proceedings of the 29th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 130–137.

Daudé, Jordi, Lluis Padro, and German Rigau (2000). "Mapping wordnets using structural information". In: *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 504–511.

Daude, Jordi, Luiss Padro, and German Rigau (2003). "Validation and tuning of Wordnet mapping techniques". In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'03)*. Borovets, Bulgaria.

Daudé, J., L. Padró, and G. Rigau (2001). "A Complete WN 1.5 to WN 1.6 mapping". In: *Proceedings of NAACL Workshop" WordNet and Other Lexical Resources: Applications, Extensions and Customizations". Pittsburg, PA*. Citeseer.

Diab, Mona (2004). "Relieving the data acquisition bottleneck in word sense disambiguation". In: *Proceedings of the 42nd annual meeting on association for computational linguistics*. Association for Computational Linguistics, p. 303.

Diab, Mona and Philip Resnik (2002). "An unsupervised method for word sense tagging using parallel corpora". In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 255–262.

Diab, Mona Talat (2003). "Word Sense Disambiguation Within a Multilingual Framework". PhD thesis. College Park, MD, USA.

Dyer, Chris, Victor Chahuneau, and Noah A. Smith (2013). "A Simple, Fast, and Effective Reparameterization of IBM Model 2." In: *HLT-NAACL*. Citeseer, pp. 644–648.

Edmonds, Philip and Scott Cotton (2001). "SENSEVAL-2: overview". In: *The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems*. Association for Computational Linguistics, pp. 1–5.

Erjavec, Tomaz (2004). "MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora." In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal: European Language Resources Association (ELRA).

Erk, Katrin and Carlo Strapparava, eds. (2010). *Proceedings of the 5th International Workshop on Semantic Evaluation*. Uppsala, Sweden: Association for Computational Linguistics.

Evert, Stefan and Andrew Hardie (2011). "Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium". In:

Fellbaum, Christiane (1998). *WordNet*. Blackwell Publishing Ltd.

Gale, William, Kenneth Ward Church, and David Yarowsky (1992a). "Estimating upper and lower bounds on the performance of word-sense disambiguation programs". In: *Proceedings of the 30th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 249–256.

Gale, William A., Kenneth W. Church, and David Yarowsky (1992b). "One Sense Per Discourse". In: *Proceedings of the Workshop on Speech and Natural Language*. HLT '91. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 233–237.

— (1992c). *Using Bilingual Materials to Develop Word Sense Disambiguation Methods*.

Gliozzo, Alfio Massimiliano, Marcello Ranieri, and Carlo Strapparava (2005). "Crossing parallel corpora and multilingual lexical databases for WSD". In: *Computational Linguistics and Intelligent Text Processing*. Springer, pp. 242–245.

Ide, Nancy (2000). "Cross-lingual sense determination: Can it work?" In: *Computers and the Humanities* 34.1-2, pp. 223–234.

Ide, Nancy and Laurent Romary (2003). "Outline of the international standard linguistic annotation framework". In: *Proceedings of the ACL 2003 workshop on Linguistic annotation: getting the model right-Volume 19*. Association for Computational Linguistics, pp. 1–5.

Ide, Nancy and Jean Véronis (1998). "Introduction to the special issue on word sense disambiguation: the state of the art". In: *Computational linguistics* 24.1, pp. 2–40.

Ide, Nancy and Yorick Wilks (2006). "Making sense about sense". In: *Word sense disambiguation*. Springer, pp. 47–73.

Ide, Nancy, Patrice Bonhomme, and Laurent Romary (2000). "An XML-based Encoding Standard for Linguistic Corpora". In: *Proceedings of the Second International Conference on Language Resources and Evaluation*, pp. 825–830.

Ide, Nancy, Tomaz Erjavec, and Dan Tufis (2002). "Sense discrimination with parallel corpora". In: *Proceedings of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions-Volume 8*. Association for Computational Linguistics, pp. 61–66.

Ion, Radu (2007). "Metode de dezambiguizare semantica automata. Aplicat ii pentru limbile englezas i romana ("Word Sense Disambiguation methods applied to English and Romanian")". PhD thesis. Bucharest: Research Institute for Artificial Intelligence (RACAI), Romanian Academy.

Ion, Radu and Dan Tufiş (2009). "Multilingual versus monolingual word sense disambiguation". en. In: *International Journal of Speech Technology* 12.2-3, pp. 113–124.

Isahara, Hitoshi et al. (2008). "Development of the Japanese WordNet." In: *LREC*.

Kilgarriff, A. and M. Palmer (2000). "Introduction to the Special Issue on SENSEVAL". In: *Computers and the Humanities* 34.1/2, pp. 1–13. ISSN: 00104817.

Kilgarriff, Adam (1997). ""I Don't Believe in Word Senses"". In: *Computers and the Humanities* 31.2, pp. 91–113.

— (1998). "Senseval: An exercise in evaluating word sense disambiguation programs". In: *Proc. of the first international conference on language resources and evaluation*, pp. 581–588.

— (2006). "Word Senses". In: *Word Sense Disambiguation: Algorithms and Applications*. Ed. by Eneko Agirre and Philip Edmonds. Dordrecht: Springer Netherlands, pp. 29–46.

Koehn, Philipp (2005). "Europarl: A parallel corpus for statistical machine translation". In: *MT summit*. Vol. 5, pp. 79–86.

Kucera, Henry and W. Nelson Francis (1982). *Frequency analysis of English usage: Lexicon and grammar*. Houghton Mifflin.

Landes, Shari, Claudia Leacock, and Christiane Fellbaum (1998). "Building semantic concordances". In: *In Fellbaum (1998), chapter 8*.

Lefever, Els (2012). "ParaSense: parallel corpora for word sense disambiguation". PhD thesis. Ghent University.

Lefever, Els and Véronique Hoste (2013). “Semeval-2013 task 10: Cross-lingual word sense disambiguation”. In: *Proc. of SemEval*, pp. 158–166.

— (2014). “Parallel corpora make sense: Bypassing the knowledge acquisition bottleneck for Word Sense Disambiguation”. en. In: *International Journal of Corpus Linguistics* 19.3, pp. 333–367.

Lefever, Els, Véronique Hoste, and Martine De Cock (2011). “Parasense or how to use parallel corpora for word sense disambiguation”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, pp. 317–322.

Lefever, Els, Véronique Hoste, and Martine De Cock (2013). “Five languages are better than one: an attempt to bypass the data acquisition bottleneck for WSD”. In: *Computational Linguistics and Intelligent Text Processing*. Springer, pp. 343–354.

Lesk, Michael (1986). “Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone”. In: *Proceedings of the 5th annual international conference on Systems documentation*. ACM, pp. 24–26.

Li, Hang and Cong Li (2004). “Word Translation Disambiguation Using Bilingual Bootstrapping”. In: *Computational Linguistics* 30.1, pp. 1–22.

Lupu, Monica, Diana Trandabat, and Maria Husarciuc (2005). “A Romanian SemCor aligned to the English and Italian MultiSemCor”. In: *1st ROMANCE FrameNet Workshop at EUROLAN*. Citeseer, pp. 20–27.

Lyons, John (1977). *Semantics*. Vol. 2. Semantics. Cambridge University Press.

Mana, Nadia and Ornella Corazzari (2002). "The Lexico-semantic Annotation of an Italian Treebank." In: *LREC*. European Language Resources Association.

Manandhar, Suresh and Deniz Yuret, eds. (2013). *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Atlanta, Georgia, USA: Association for Computational Linguistics.

Markert, Katja and Malvina Nissim (2007). "Semeval-2007 task 08: Metonymy resolution at semeval-2007". In: *Proceedings of the 4th International Workshop on Semantic Evaluations*. Association for Computational Linguistics, pp. 36–41.

McCarthy, Diana and John Carroll (2003). "Disambiguating nouns, verbs, and adjectives using automatically acquired selectional preferences". In: *Computational Linguistics* 29.4, pp. 639–654.

Mihalcea, Rada and Phil Edmonds (2004). "Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text". In: *Barcelona, Spain, Association for Computational Linguistics Conference (ACL2004)*, pp. 44–48.

Mihalcea, Rada and Dan I. Moldovan (1999). "Automatic method for generating sense tagged corpora". In: *Proceedings of the National Conference on Artificial Intelligence*. AAAI, pp. 461–466.

Miller, George A. and Christiane Fellbaum (1991). "Semantic networks of english". In: *Cognition* 41.1–3, pp. 197–229.

Miller, George A et al. (1990). "Introduction to WordNet: An on-line lexical database". In: *International journal of lexicography* 3.4, pp. 235–244.

Mititelu, Verginica Barbu, Ștefan Daniel Dumitrescu, and Dan Tufiș (2014). "News about the Romanian Wordnet". In: *Proceedings of the Seventh Global Wordnet Conference, GWC*, pp. 268–275.

Morato, Jorge et al. (2004). "Wordnet applications". In: *Global Wordnet Conference*. Vol. 2, pp. 270–278.

Murphy, M. Lynne (2010). *Lexical Meaning*. Cambridge Books Online. Cambridge University Press. ISBN: 9780511780684.

Nakov, Preslav and Torsten Zesch, eds. (2014). *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Dublin, Ireland: Association for Computational Linguistics and Dublin City University.

Nakov, Preslav et al., eds. (2015). *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Denver, Colorado: Association for Computational Linguistics.

Navigli, Roberto (2006). "Meaningful clustering of senses helps boost word sense disambiguation performance". In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 105–112.

— (2009). "Word sense disambiguation: A survey". In: *ACM Computing Surveys* 41.2, pp. 1–69.

Navigli, Roberto and Mirella Lapata (2007). "Graph Connectivity Measures for Unsupervised Word Sense Disambiguation". In: *Proceedings of the 20th International Joint Conference on Artificial Intelligence*. Hyderabad, India, pp. 1683–1688.

Navigli, Roberto and Simone Paolo Ponzetto (2010). "BabelNet: Building a very large multilingual semantic network". In: *Proceedings of the 48th annual meet-*

ing of the association for computational linguistics. Association for Computational Linguistics, pp. 216–225.

Ng, Hwee Tou (1997). “Getting serious about word sense disambiguation”. In: *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How*, pp. 1–7.

Ng, Hwee Tou and Hian Beng Lee (1996). “Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-based Approach”. In: *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*. ACL ’96. Santa Cruz, California: Association for Computational Linguistics, pp. 40–47.

Ng, Hwee Tou, Bin Wang, and Yee Seng Chan (2003). “Exploiting parallel texts for word sense disambiguation: An empirical study”. In: *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, pp. 455–462.

Niles, Ian and Adam Pease (2001). “Towards a standard upper ontology”. In: *Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001*. ACM, pp. 2–9.

Och, Franz Josef (2002). “Statistical machine translation: from single-word models to alignment templates”. PhD thesis. Bibliothek der RWTH Aachen.

Palmer, Martha, Hwee Tou Ng, and Hoa Trang Dang (2006). “Evaluation of WSD Systems”. In: *Word Sense Disambiguation: Algorithms and Applications*. Ed. by Eneko Agirre and Philip Edmonds. Dordrecht: Springer Netherlands, pp. 75–106. ISBN: 978-1-4020-4809-8.

Peters, Wim et al. (1998). “Cross-linguistic alignment of wordnets with an interlingual-index”. In: *EuroWordNet: A multilingual database with lexical semantic networks*. Springer, pp. 149–179.

Petrolito, Tommaso (2016). “A language-independent LESK based approach to Word Sense Disambiguation.” In: *Proceedings of the Eighth Global WordNet Conference*. Bucharest, pp. 273–279.

Petrov, Slav, Dipanjan Das, and Ryan McDonald (2012). “A Universal Part-of-Speech Tagset”. In: *LREC*. European Language Resources Association (ELRA), pp. 2089–2096.

Pianta, Emanuele and Luisa Bentivogli (2004). “Knowledge intensive word alignment with KNOWA”. In: *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, p. 1086.

Pianta, Emanuele, Luisa Bentivogli, and Christian Girardi (2002). “MultiWordNet: Developing an Aligned Multilingual Database”. In: *In Proceedings of the First International Conference on Global WordNet*. Mysore, India, pp. 293–302.

Resnik, Philip (1995). “Disambiguating noun groupings with respect to WordNet senses”. In: *arXiv preprint cmp-lg/9511006*.

Resnik, Philip and David Yarowsky (1997). “A Perspective on Word Sense Disambiguation Methods and Their Evaluation”. In: pp. 79–86.

— (1999). “Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation”. In: *Natural language engineering* 5.02, pp. 113–133.

Roventini, Adriana et al. (2000). “ItalWordNet: a Large Semantic Database for Italian”. In: *LREC*. European Language Resources Association.

Sanderson, Mark (1994). “Word sense disambiguation and information retrieval”. In: *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*. Springer-Verlag New York, Inc., pp. 142–151.

Schutze, Hinrich and Jan O. Pedersen (1995). “Information Retrieval Based on Word Senses”. In: *In Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval*, pp. 161–175.

Sinha, Ravi and Rada Mihalcea (2007). “Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity”. In: *Proceedings of the International Conference on Semantic Computing*. ICSC ’07. Washington, DC, USA: IEEE Computer Society, pp. 363–369. isbn: 0-7695-2997-6.

Stamou Sofia, Kemal Oflazer Karel Pala Dimitris Christoudoulakis Dan Cristea Dan Tufis Svetla Koeva George Totkov Dominique Dutoit and Maria Grigoriadou (2002). “Balkanet: A multilingual semantic network for the balkan languages”. In: *Proceedings of the International Wordnet Conference*. Mysore, India, pp. 21–25.

Steinberger, Ralf et al. (2013). “DGT-TM: A freely Available Translation Memory in 22 Languages”. In: *CoRR* abs/1309.5226.

Tan, Liling and Francis Bond (2014). “NTU-MC toolkit: Annotating a linguistically diverse corpus”. In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pp. 86–89.

Tufiş, Dan (1999). “Tiered Tagging and Combined Language Models Classifiers”. In: *Text, Speech and Dialogue (TSD 1999)*. Ed. by Václav Matousek et al. Lecture Notes in Artificial Intelligence 1692. Springer Berlin / Heidelberg, pp. 28–33.

— (2005). “Word Sense Disambiguation: A Case Study on the Granularity of Sense Distinctions”. In: *WSEAS Transactions on Information Science and Applications*. Ed. by S. Bojkovic Zoran. Vol. 2. 2, pp. 183–188.

— (2006). “From Word Alignment to Word Senses, via Multilingual Wordnets”. In: *Computer Science Journal of Moldova* 14.1(40), pp. 3–33.

Tufiş, Dan and Nancy Ide (2004). “Word sense disambiguation as a wordnets validation method in Balkanet”. In: *Proceedings of the 4th LREC Conference*. Lisbona, pp. 741–744.

Tufiş, Dan and Radu Ion (2004). “Multilingual Word Sense Disambiguation Using Aligned Wordnets”. In: *Romanian Journal on Information Science and Technology* 7.2-3. Ed. by Dan Tufiş. Special Issue on BalkaNet, pp. 198–214.

Tufiş, Dan et al. (2004). “The Romanian Wordnet”. In: *Romanian Journal on Information Science and Technology* 7.2-3. Ed. by Dan Tufiş. Special Issue on BalkaNet, pp. 105–122.

Tufiş, Dan, Radu Ion, and Nancy Ide (2004). “Fine-grained word sense disambiguation based on parallel corpora, word alignment, word clustering and aligned wordnets”. In: *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, p. 1312.

Tufiş, Dan et al. (2008). “Romanian WordNet: Current state, new applications and prospects”. In: *Proceedings of 4th Global WordNet Conference, GWC*, pp. 441–452.

Tufiş, Dan et al. (2013). “The Romanian wordnet in a nutshell”. In: *Language Resources and Evaluation* 47.4, pp. 1305–1314.

Tufiș, Dan and Radu Ion (2003). "Word sense clustering based on translation equivalence in parallel texts; a case study in Romanian". In: *Proceedings of the International Conference on Speech and Dialog–SPED*, pp. 13–26.

Turing, Alan M (1950). "Computing machinery and intelligence". In: *Mind* 59.236, pp. 433–460.

Utiyama, Masao and Hitoshi Isahara (2003). "Reliable measures for aligning Japanese-English news articles and sentences". In: *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, pp. 72–79.

Véronis, Jean and Philippe Langlais (2000). "Evaluation of parallel text alignment systems". In: *Parallel text processing*. Springer, pp. 369–388.

Vossen, Piek (1998). "EuroWordNet: building a multilingual database with wordnets for European languages". In: *The ELRA Newsletter* 3.1, pp. 7–10.

— (2002). "WordNet, EuroWordNet and Global WordNet". In: *Revue française de linguistique appliquée* 7.1, pp. 27–38.

Vossen, Piek, Wim Peters, and Julio Gonzalo (1999). "Towards a universal index of meaning". In: *SIGLEX99: Standardizing Lexical Resources*.

Wasow, Thomas, Amy Perfors, and David Beaver (2005). "The puzzle of ambiguity". In: *Morphology and the web of grammar: Essays in memory of Steven G. Lapointe*, pp. 265–282.

Weaver, Warren (1949). "Translation". In: *Machine Translation of Languages*. Ed. by William N. Locke and A. Donald Boothe. Reprinted from a memorandum written by Weaver in 1949. Cambridge, MA: MIT Press, pp. 15–23.

Winston, Morton E., Roger Chaffin, and Douglas Herrmann (1987). “A taxonomy of part-whole relations”. In: *Cognitive science* 11.4, pp. 417–444.

Yarowsky, David (1995). “Unsupervised word sense disambiguation rivaling supervised methods”. In: *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 189–196.

Zipf, George Kingsley (1949). *Human behavior and the principle of least effort.* Addison-Wesley press.