Multilingual Computational Semantic Lexicons in Action: The WYSINNWYG Approach to NLP

Evelyne Viegas
New Mexico State University
Computing Research Laboratory
Las Cruces, NM 88003
USA
viegas@crl.nmsu.edu

Abstract

Much effort has been put into computational lexicons over the years, and most systems give much room to (lexical) semantic data. However, in these systems, the effort put on the study and representation of lexical items to express the underlying continuum existing in 1) language vagueness and polysemy, and 2) language gaps and mismatches, has remained embryonic. A sense enumeration approach fails from a theoretical point of view to capture the core meaning of words, let alone relate word meanings to one another, and complicates the task of NLP by multiplying ambiguities in analysis and choices in generation. In this paper, I study computational semantic lexicon representation from a multilingual point of view, reconciling different approaches to lexicon representation: i) vagueness for lexemes which have a more or less finer grained semantics with respect to other languages; ii) underspecification for lexemes which have multiple related facets; and, iii) lexical rules to relate systematic polysemy to systematic ambiguity. I build on a What You See Is Not Necessarily What You Get (WYSINNWYG) approach to provide the NLP system with the “right” lexical data already tuned towards a particular task. In order to do so, I argue for a lexical semantic approach to lexicon representation. I exemplify my study through a cross-linguistic investigation on spatially-based expressions.

1 A Cross-linguistic Investigation on Spatially-based Expressions

In this paper, I argue for computational semantic lexicons as active knowledge sources in order to provide Natural Language Processing (NLP) systems with the “right” lexical semantic representation to accomplish a particular task. In other words, lexicon entries are “pre-digested”, via a lexical processor, to best fit an NLP task. This What You See (in your lexicon) Is Not Necessarily What You Get (as input to your program) (WYSINNWYG) approach requires the adoption of a symbolic paradigm. Formally, I use a combination of three different approaches to lexicon representations: (1) lexico-semantic vagueness, for lexemes which have a more or less finer grained semantics with respect to other languages (for instance en in Spanish is vague between the Contact and Container senses of the Location, whereas in English it is finer grained, with on for the former and in for the latter); (2) lexico-semantic underspecification, for lexemes which have multiple related facets (such as for instance, door which is underspecified with respect to its Aperture or PhysicalObject meanings); and, (3) lexical rules, to relate systematic polysemy to systematic ambiguity (such as the Food Or Animal rule for lamb).

I illustrate the WYSINNWYG approach via a cross-linguistic investigation (English, French, Spanish) on spatially-based expressions, as lexicalised, for instance, in the prepositions in, above, on, ..., verbs traverser, (“go” across) in French, predicative nouns montée, (going up) in French, or in adjectives upright. Processing spatially-based expressions in a multilingual environment is a difficult problem as these lexemes exhibit a high degree of polysemy (in particular for prepositions) and of language gaps (i.e., when there is not a one-to-one mapping between languages, whatever the linguistic level; lexical, semantic, syntactic, etc). Therefore, processing these expressions or words in a multilingual environment minimally involves having a solution for treating: (a) syntactic divergences, swim across → traverser ... à la nage in French (cross ... swimming); (b) semantic mismatches, river translates into fleuve, rivière in French; and (c), cases which lie in between clear-cut cases of language gaps (stand → se tenir debout/se tenir, lie → se tenir allongé/se tenir). Researchers have dealt with a) and/or b), whereas WYSINNWYG presents a uniform treatment of a), b) and c), by allowing words to have their meanings vary in context.

In this paper, I restrict my cross-linguistic study to the (lexical) semantics of words with a focus on spatially-based expressions, and consider literal or non-figurative meanings only. In the next sections, I address representational problems which must be solved in order to best capture the phenom-
ena of ambiguity, polysemy and language gaps from a lexical semantic viewpoint. I then present three different ways of capturing the phenomena: lexico-semantic vagueness, lexico-semantic underspecification and lexical rules.

1.1 The Language Gap Problem
Upon a close examination of empirical data, it is often difficult to classify a translation pair as a syntactic divergence (e.g., Dorr, 1990; Levin and Nirenburg, 1993), as in

he limped up the stairs → il monta les marches en boitant (French) (he went up the stairs limping) or a semantic mismatch (e.g., Palmer and Zhibiao, 1995; Kameyama et al., 1991), as in lie, stand → se tenir (French). Moreover, lie and stand could be translated as se tenir couché/allongé (be lying) and se tenir debout (be up) respectively, thus presenting a case of divergence, or they could both be translated into French as se tenir, thus presenting a case of conflation, (Talmy, 1985). Depending on the semantics of the first argument, one might want to generate the divergence, (e.g., se tenir debout/couché), or not (e.g., se tenir), thus considering se tenir as a mismatch as in (1):

(1) Pablo se tenait au milieu de la chambre.
(Sartre)
(Pablo stood in the middle of the bedroom.)

In order to account for all these language variations, one cannot “freeze” the meanings of language pairs. In section 2.1, I show that by adopting a continuum perspective, that is using a knowledge-based approach where I make the distinction between lexical and semantic knowledge, cases in between syntactic divergences and semantic mismatches (se tenir) can be accounted for in a uniform way. Practically, the proposed method can be applied to interlingua approaches and transfer approaches, when these latter encode a layer of semantic information.

1.2 The Lexicon Representation Problem
Within the paradigm of knowledge-based approaches, there are still lexicon representation issues to be addressed in order to treat these language gaps. It has been well documented in the literature of this past decade that a sense enumeration approach fails from a theoretical point of view to capture the core meaning of words (e.g., (Ostler and Atkins, 1992), (Boguraev and Pustejovsky, 1990),...) and complicates from a practical viewpoint the task of NLP by multiplying ambiguities in analysis and choices in generation.

Within Machine Translation (MT), this approach has led researchers to “add” ambiguity in a language which did not have it from a monolingual perspective. Ambiguity is added at the lexical level within transfer based approaches (“rivière” → RIVER - DESTINATION: RIVER; “fleuve” → RIVER - DESTINATION: SEA; “river” → RIVER - DESTINATION: SEA, RIVER), whereas again “river” in English is not ambiguous with respect to its destination.

In this paper, I show that ambiguity can be minimised if one stops considering knowledge sources as “static” ones in order to consider them as active ones instead. More specifically, I show that building on a computational theory of lexico-semantic vagueness and underspecification which merges computational concerns with theoretical concerns enables an NLP system to cope with polysemy and language gaps in a more effective way.

Let us consider the following simplified input semantics (IS):

(2) PositionState(Theme:Plate,Location:Table),

This can be generated in Spanish as El plato esta en la mesa; where Location is lexicalised as en in Figure 1.

To generate (2) into English, requires the system to further specify Location for English as LocContact, in order to generate The plate is on the table, where on1 corresponds to the Spanish en1, sub-sense of en, as shown in Figure 1.

[Figure 1: Subset of the Semantic Types for Prepositions]

From a monolingual perspective, there is no need to differentiate in Spanish between the 3 types of Location as LocContact, LocContainer and LocBuilding, as these distinctions are irrelevant for Span-
ish analysis or generation, with respect to Figure 1. However, within a multilingual framework, it becomes necessary to further distinguish Location, in order to generate English from (2). In the next sections, I will show that lexical semantic hierarchies are better suited to account for polysemous lexemes than lexical or semantic hierarchies alone, for multilingual (and monolingual) processing.

2 The WYSINNWYG Approach

I argue that treating lexical ambiguity or polysemy and language gaps computationally requires 1) fine-grained lexical semantic type hierarchies, and 2) to allow words to have their meanings vary in context. Much effort has been put into lexicons over the years, and most systems give more room to lexical data. However, most approaches to lexicon representation in NLP systems have been motivated more by computational concerns (economy, efficiency) than by the desire for a computational linguistic account, where the concern of explaining a phenomenon is as important as pure computational concerns. In this paper, I adopt a computational linguistic perspective, showing however, how these representations are best fitted to serve knowledge-driven NLP systems.

2.1 A Continuum Perspective on Language Gaps

I argue that resolving language gaps (divergences, mismatches, and cases in between) is a generation issue and minimally involves:

1) using a knowledge-based approach to represent the lexical semantics of lexemes;
2) developing a computational theory of lexical-semantic vagueness, underspecification, and lexical rules;

In this paper, I only address lexical representational issues, leaving the generation issues (such as the use of planning techniques, the integration of the process in lexical choice) aside.¹

I illustrate through some examples below, how a compositional semantics approach, e.g. knowledge-based, can help in dealing with language gaps.² I will use the French (se tenir) and English (stand, lie) simplified entries below, in my illustration of mismatches between the generator and the lexicons. Semantic types are coded in the sense feature:

Figure 2 illustrates a subset of the Semantic Type Hierarchy (STH) common to all dictionaries and of two subsets of the Lexical Type Hierarchy (LTH) for French and English.

I illustrate below three main types of gaps between the input semantics (IS) to the generator and the lexicon entries (LEX) of the language in which to generate. I focus on the generation of the predicate:

(i) IS - LEX exact match Generating, in French, from the simplified IS below (3),

(3) PositionState(agent:john, against:wall)

is easy as there is a single French word in (3) that lexicalises the concept PositionState, which is se tenir. Therefore se tenir is generated in John se tenait contre le mur (John was/(stood) against the wall).

Figure 2: Example of an STH linked to a Fragment of the French and English LTHs.

I illustrate below three main types of gaps between the input semantics (IS) to the generator and the lexicon entries (LEX) of the language in which to generate. I focus on the generation of the predicate:

(i) IS - LEX exact match Generating, in French, from the simplified IS below (3),

(3) PositionState(agent:john, against:wall)
(ii) IS - LEX vagueness Generating, in French, from the partial IS below (4),

(4) PsVertical(agent: john, against: wall)

needs extra work from the generator, with respect to the lexicon entry for French. In Figure 2, one can see in STH that PsVertical is a sub-type of PositionState, which has a mapping in LTH for French to se-tenir3. This illustrates a case of vagueness between English and French. In this case, the generator will generate the same sentence John se tenait contre le mur, as is the case for the exact match in (i). Note that generating the divergence se tenait debout (stand upright) although correct and grammatical, would emphasise the position of John which was not necessarily focused in (4). The divergence can be generated by "composing" PsVertical as PositionState (lexicalised as se tenir) and Vertical (lexicalised as debout).

(iii) IS - LEX Underspecification Generating, in French, from the partial IS below (5),

(5) PsVertical(agent: john, against: wall, time: t1) & PsHorizontal(agent: john, against: wall, time: t2) & t1 < t2

needs extra work from the lexicon processor, with respect to the entries presented here, as one does not want to end up generating John se tint contre le mur puis il se tint contre le mur (John was against the wall then he was against the wall). Because of the conjunctions here, one cannot just consider se tenir as vague with respect to lie and stand. This illustrates a lexicon in action, where the lexical processor must process se tenir as underspecified:

PositionState → PsVertical ∨ PsHorizontal

The lexical processor will thus produce the divergences se tenir debout (stand) and se tenir allongé (lying) to generate (with some generation processing such as lexical choice, ellipsis, pronominalisation, etc) John se tenait (debout) contre le mur puis s'allongea contre lui (John was standing against the wall then he lied against it).

Where the continuum perspective comes in, is that we do not want to "freeze" the meanings of words once and for all. As we just saw, in French one might want to generate se tenir debout or just se tenir depending on the semantics of its arguments and also depending on the context as in (5).

In the WYSINNWYG approach, words are allowed to have their "meanings" vary in context. In other words, the literal meaning(s) coded in the lexicon is/are the "closest" possible meaning(s) of a word within the STH context, and by enriching the discourse context (dc), one ends up "specialising" or "generalising" the meaning(s) of the word, using formally two hierarchies: semantic (STH) and lexical (LTH), enabling different types of lexicon representations: vagueness, underspecification and lexical rules.

2.2 A Truly Multilingual Hierarchy

Multilingual lexicons are usually monolingual lexicons connected via translation links (Tlinks), whereas truly multilingual lexicons, as defined by (Cahill and Gazdar, 1995), involve n + 1 hierarchies, thus involving an additional abstract hierarchy containing information shared by two or more languages. Figure 3 illustrates the STH which is shared by all lexicons (French, English, Spanish, etc), and the lexical MLTH which involves the abstract hierarchy shared by all LTHs.

The lexicons themselves are also organised as language lexical type hierarchies (Spanish LTH, English LTH in Figure 3). For instance, the English dictionary (eng-lexeme) has the English prepositions (eng-prep) as one of its sub-types, which itself has as subtypes all the English prepositions (along, through, on, in, ...). These prepositions have in turn subtypes (for instance, on has on1, on2, ...), which can themselves have subtypes (on11, on12, ...). All these language dependent LTHs inherit part of their information from a truly Multilingual Lexical Type Hi-
erarchy (MLTH), which contains information shared by all lexicons. There might be several levels of sharing, for instance, family-related languages sharing.

Lexical types are linked to the STH via their language LTH and the MLTH, so that these lexicons can be used by either monolingual or multilingual processing. The advantages of a MTLH extend to 1) lexicon acquisition, by allowing lexicons to inherit information from the abstract level hierarchy. This is even more useful when acquiring family-related languages; and 2) robustness, as the lexical processors can try to “make guesses” on the assignment of a sense to a lexeme absent from a dictionary, based on similarities in morphology or orthography, with other family-related language lexemes. 31

2.3 Vagueness, Underspecification and Lexical Rules

The STH along with the LTH allow the lexicographers to leave the meaning of some lexemes as vague or underspecified. The vagueness or underspecification typing allows the lexical processor to specialise or generalise the meaning of a lexeme, for a particular task and on a needed basis. Formally, generalisation and specialisation can be done in various ways, as specified for instance in (Kameyama et al., 1991), (Poesio, 1996), (Mahesh et al., 1997).

2.3.1 Lexicon Vagueness

A lexicon entry is considered as vague when its semantics is typed using a general monomorphic type covering multiple senses, as is the case of the French entry “se-tenir3”, or the Spanish preposition “por” in (8), and typical examples in lexical semantics, such as door which is typed as PHYSICAL_OBJECT-OR-APERTURE. 5

In other words, lex is vague, if m is in a type/sub-type relation with all its immediate sub-types.

2.3.2 Lexicon Underspecification

The meaning of a lexeme is considered as underspecified when its semantics is represented via a polymorphic type, which presents a disjunction of semantic types, thus covering different polysemous senses, as is the case of the Spanish preposition “por” in (8), and typical examples in lexical semantics, such as door which is typed as PHYSICAL_OBJECT-OR-APERTURE. 5

It is at processing time only, and on a needed basis only, that the semantic type Through-OR-Along for por can be further processed as either Through, or Along, ..., thus allowing the generator or analyser to find the appropriate representation depending on the task. Disambiguating “por” to generate English, requires that the lexeme be embedded within the discourse context, where the filled arguments of the prepositions will provide semantic information under constraints. For instance, walk and river could contribute to the disambiguation of por as Along.

Lexicon underspecification is represented by mapping lex (the citation form of a word x) to a semantic polymorphic type p, which belongs to STH, then:

(9) x is underspecified iff

\[ \exists p \in STH : p = \text{MAPS}(dc, lex(x)) \land \exists s \subseteq STH : p = vs \land \text{Card}(s) \geq 2 \]

In other words, lex is underspecified, if p is a disjunction of types, and no type/sub-type relation is required.

31 I have not investigated this issue yet, but see (Cahill, 1998) for promising results with respect to making guesses on phonology.

3 See (Sanfilippo, 1998) and (Buitelaar, 1997) for different computational treatments of underspecified representations. The former deals with multiple subcategorisations (whereas I am also interested in polysemous senses), the latter includes homonyms, which I agree with Pinkal (1995) should be left apart.

5 I believe that lexicosemantic underspecification is concerned with polysemous lexemes only (such as door, book, etc) and not homonyms (such as bank as financial-bank or river-bank) called H-Type ambiguous in (Pinkal, 1995). I believe the H-Type ambiguous lexemes should be related via their lexical form only, while their semantic types should remain unrelated, i.e., there is no needs to introduce a “disjunction fallacy” as in (Poesio, 1996). It might be the case that homonyms require pragmatic underspecification as suggested, for instance, in (Nunberg, 1979), but in any case are beyond the scope of this paper.
2.4 Lexical Rules

Lexical rules (LRs) are used in WYSINNWYG to relate systematic ambiguity to systematic polysemy. They seem more appropriate than underspecification for relating the meanings of lexemes such as "lamb" or "haddock" which can be either of type Animal or Food (Pustejovsky, 1995, pp. 224). LRs and their application time in NLP have received a lot of attention (e.g., Copestake and Briscoe, 1996; Viegas et al., 1996), therefore, I will not develop them further in this paper, as the rules themselves activated by the lexical processor produce different entries, with neither type/sub-type relations nor disjunction between the semantic types of the old and new entries. In WYSINNWYG, lexicon entries related via LRs are neither vague nor underspecified. For instance, the "grinding rule" of Copestake and Briscoe for linking the systematic Animal - Food polysemy as in mutton/sheep or in French where we have a conflation in mouton, allows us to link the entries in English and sub-senses in French, without having to cope with the semantic "disjunction fallacy problem" of (Poesio, 1996).

3 Conclusions - Perspectives

I have argued for active knowledge sources within a knowledge-based approach, so that lexicon entries can be processed to best fit a particular NLP task. I adopted a computational linguistic perspective in order to explain language phenomena such as language gaps and polysemy. I argued for semantic and lexical type hierarchies. The former is shared by all dictionaries, whereas the latter can be organised as a truly multilingual hierarchy. In that respect, this work differs from (Han et al., 1996) in that I do not suggest an ontology per language, but argue on the contrary for one semantic hierarchy shared by all dictionaries. Other works which have dealt with mismatches, e.g., (Dorr and Voss, 1998) with their interlingua and knowledge representations, (Sérasset, 1994) with his "interlingua acceptations", or (Kameyama, et al, 1991) with their inbons, cannot account for cases which lie in between clear-cut cases of divergences and mismatches such as the example "se tenir" discussed in this paper.

I have shown that enabling lexicon entries to be typed as either lexically vague or underspecified, or linked via LRs, allows us to account for the variations of word meanings in different discourse contexts. Most of the works in computational lexical semantics have dealt with either underspecification or LRs, trying to favour one representation over the other. There was previously no computational treat-

---

6However, I do not preclude that there might be different views on the semantic hierarchy depending on the languages considered: "filters" could be applied to the STH to only show the relevant parts of it for some family-related languages.

7The fact that some cultures eat "living" creatures would require to type these lexemes using underspecification (food-OR-animal) instead of a lexical rule in their cultures.
tenir" → PositionState) and use mechanisms (such as generalisation, specialisation) to modulate their meanings in context (e.g., "se tenir" → PsVertical). In other words, WYSINNWYG helps not only in sense selection but also in sense modulation.

Further research involves investigating representation formalisms, as discussed in (Briscoe et al., 1993) to best implement these type inheritance hierarchies.

4 Acknowledgements

This work has been supported in part by DoD under contract number MDA-904-92-C-5189. I would like to thank my colleagues at CRL for commenting on a former version of this paper. I am also grateful to John Barnden, Pierrette Bouillon, Boyan Onyshkevych, Martha Palmer, and the anonymous reviewers for their useful comments.

References

L. Talmy. 1985. Lexicalization Patterns: semantic structure in lexical forms. In Shopen (ed), Language Typology and Syntactic Description III. CUP.