The semantic and stylistic differentiation of synonyms and near-synonyms

Chrysanne DiMarco
Department of Computer Science
University of Waterloo
Waterloo, Ontario
Canada N2L 3G1
cdimarco@logos.uwaterloo.ca

Graeme Hirst and Manfred Stede
Department of Computer Science
University of Toronto
Toronto, Ontario
Canada M5S 1A4
gh and mstede@cs.toronto.edu

1 Introduction

If we want to describe the action of someone who is looking out a window for an extended time, how do we choose between the words gazing, staring, and peering? What exactly is the difference between an argument, a dispute, and a row? In this paper, we describe our research in progress on the problem of lexical choice and the representations of world knowledge and of lexical structure and meaning that the task requires. In particular, we wish to deal with nuances and subtleties of denotation and connotation—shades of meaning and of style—such as those illustrated by the examples above.

We are studying the task in two related contexts: machine translation, and the generation of multilingual text from a single representation of content. This work brings together several elements of our earlier research: unilingual lexical choice (Miezis 1988); multilingual generation (Rösner and Stede 1992a,b); representing and preserving stylistic nuances in translation (DiMarco 1990; DiMarco and Hirst 1990; Mah 1991); and, more generally, analyzing and generating stylistic nuances in text (DiMarco and Hirst 1993; DiMarco et al 1992; Makuta-Giluk 1991; Makuta-Giluk and DiMarco 1993; BenHas-sine 1992; Green 1992a,b, 1993; Hoyt forthcoming).

In the present paper, we concentrate on issues in lexical representation. We describe a methodology, based on dictionary usage notes, that we are using to discover the dimensions along which similar words can be differentiated, and we discuss a two-part representation for lexical differentiation. (Our related work on lexical choice itself and its integration with other components of text generation is discussed by Stede (1993a,b, forthcoming).)

2 Synonymy and plesionymy within and across languages

While absolute synonymy—the interchangeability of pairs of words in any context—is rare at best, it is common to find pairs or sets of words (or, more strictly, word senses) that are synonymous to the extent that they have the same denotation, while differing in other aspects of their usage.\footnote{Cruse (1986) calls such words cognitive synonyms but we will avoid this confusing term.} Such differences can include the collocational constraints of the words (e.g., groundhog and woodchuck denote the same set of animals; yet Groundhog Day, *Woodchuck Day*) and the stylistic and interpersonal connotations of the words (e.g., die, pass away, suffocate, slim, skinny, police officer, cop, pig). In addition, many groups of words are plesionymous (Cruse 1986)—that is, nearly synonymous; forest and woods, for example, or stared and gazed, or the German words eisenschrauben, festschrauben, and festziehen.\footnote{Eisenschauben means ‘to fasten a threaded joint’, e.g., a nut on a bolt; festschrauben means ‘to fasten a threaded joint tightly’; and festziehen means ‘to fasten a threaded joint tightly with a tool’ (whereas festschrauben permits, e.g., the use of the fingers).}

The notions of synonymy and plesionymy can be made more precise by means of a notion of semantic distance (such as that invoked by Hirst (1987), for example, in lexical disambiguation); but this is troublesome to formalize satisfactorily. In this paper it will suffice to rely on an intuitive understanding.

We consider two dimensions along which words can vary: semantic and stylistic, or, equivalently, denotive and connotative. If two words differ semantically (e.g., mist, fog), then substituting one for the other in a sentence or discourse will not necessarily preserve truth conditions; the denotations are not identical. If two words differ (solely) in stylistic features (e.g., frugal, stingy), then internal substitution does preserve truth conditions, but the connotation—the stylistic and interpersonal effect of the sentence—is changed.\footnote{Recall the party game, sometimes known as “Irregular Verbs”, whose goal is to find triples of words or phrases that mean the same but vary from favorable to pejorative. Example: “I’m a renaissance person, you’re ecletic, he’s unfocused.”} Many of the semantic distinctions between plesionymous do not lend themselves to neat, taxonomic differentiation; rather, they are fuzzy, with plesionymous often having an area of overlap. For example, the boundary between forest and wood ‘tract of trees’ is vague, and there are some situations in which either word might be equally appropriate.\footnote{Observe all the hedges and degree words in this attempt to differentiate the two: “A wood” is smaller than a forest; is not so...}
and stylistic features, so intersubstitution changes both 
meaning and style. Consider:

(1) I made {an error | a blunder} in introducing her 
to my husband.

(2) The police {questioned the witnesses | interro-
gated the suspect} for many hours.

Semantically, the word blunder in (1) suggests a greater 
level of negligence than error (OALD); in addition, it is 
stylistically both more forceful and more concrete. In 
(2), the word interrogate, unlike question, suggests a 
more adversarial situation (cf DLOCE, Cornog 1992), 
and, in addition, it is a somewhat more formal word.

However, the border between denotation and connotation 
is somewhat fuzzy. For example:

(3) He {arranged | organized} the books on the 
shelves.

(4) The old professors had been {enemies | foes} for 
years.

Both choices in (3) mean ‘to put things into their proper 
place’, but arrange emphasizes the correctness or pleas-
ingness of the scheme, while organize emphasizes its 
completeness or functionality (OALD, Cornog 1992).
In (4), enemy stresses antagonism or hatred between 
the parties, whereas foe stresses active fighting rather 
than emotional reaction (Cornog 1992). Variations in 
emphasis such as these seem to sit on the boundary be-
tween variation in denotation and variation in connotation; in (3) inter-substitution seems to preserve truth 
conditions—the two forms of the sentence could describe 
the exact same situation—but this need not be true in 
general: the arrangement might be incomplete, or the 
organization not pleasing.

We can generalize these ideas across languages. A set 
of word senses drawn from two or more languages can be 
also thought of as synonymous or plesionymous if they 
meet the requisite conditions. For example, the English 
word bear ‘ursine mammal’ and the German Bär are 
synonyms. The English word soup subsumes both the 
French words soupe ‘chunky soup’ and potage ‘sieved or 
puréed soup’ (Hervey and Higgins 1992; but see foot-
note 7 below). But forest and Wald are plesionymous, 
as Wald can denote a smaller group of trees than for-
est can, for the cognate distinction between forest and 
wood in English and Wald and Holz in German breaks 
at a different point in each language; a Wald in German 
might be only a wood in English. Dutch has three 
words, hout, bos, and woud, with the first breakpoint at 
primitive, and is usually nearer to civilization. This means 
that a ‘forest’ is fairly extensive, is to some extent wild, and on the 
whole not near large towns or cities. In addition, a ‘forest’ often 
has game or wild animals in it, which a ‘wood’ does not, apart 
from the standard quota of regular rural denizens such as rabbits, 
foxes and birds of various kinds . . . ’ (Room 1985, p. 270).

<table>
<thead>
<tr>
<th>træ</th>
<th>Baum</th>
<th>boom</th>
<th>arbre</th>
<th>tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holz</td>
<td>hout</td>
<td>bois</td>
<td>wood</td>
<td></td>
</tr>
<tr>
<td>skov</td>
<td>bos</td>
<td>forêt</td>
<td>forest</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1: The relationship between words for tracts of 
trees in Danish, German, Dutch, French, and English. 
The style of the diagram and the Danish, German, and 
French columns are from Hjelmslev (1943/1961). The 
placement of the division between træ and skov is be-
cause træ also denotes wood as material, whereas in the 
other four languages, the second word in each column 
has this ambiguity.

the same place as the German and the second at the 
same place as the English and French (Henry Schogt, 
personal communication). Danish covers all situations 
with skov. (See figure 1.)

Our task is to determine and represent the differences 
between synonyms and near-synonyms, both across and 
within languages. That is, we want to describe the lex-
ical knowledge that is required to decide, in analysis, 
the exact semantic and stylistic intent of a writer’s or 
speaker’s use of a particular word, and, in generation, 
which word most precisely matches the style and mean-
ing that is to be conveyed. In translation, the problem 
arises, of course, that the target language might offer 
no single word corresponding to the exact specifications 
of the source language text; or there might be several 
words differing in style, emphasis, shade of meaning, or 
collocational requirements, from which a choice must be 
made. A similar problem occurs in text generation, es-
pecially in the generation of parallel multilingual texts.

3 The limitations of role-filling 
and selectional restrictions

We first consider a simple approach and its limitations. 
Sometimes, the distinction between a pair of plesionyms 
is clear just from their meanings, in the different require-
ments that they place on the fillers of their associated 
roles. For example, patch ‘to mend a hole in something 
by fastening a new piece of material over it’ can take a 
variety of objects (or holes therein): clothes, pipes, road 
surfaces, and so on. On the other hand, darn ‘to mend 
a hole in fabric by recreating the weave’ requires fabric 
(or a hole therein) as its object, and this is so solely 
because of its meaning; one cannot darn a hole in the
Sometimes, the selectional restrictions of a word go beyond the logical requirements of its semantics. For example, the French *réparer* 'to mend' can be used with machines, shoes, or elements of a house, but not, in modern French, for clothing or fabric (although it was so used in older French) (Anne Marie Miraglia, personal communication). (This is not merely a collocational restriction, for the class of acceptable objects is defined semantically, not lexically.) Similarly, the English *pass away* 'die' may be used only of people (or anthropomorphized pets), not plants or animals: *Many trees passed away in the drought.*

Like collocational restrictions, conceptually based role-filling and selectional restrictions are straightforward to describe in lexical entries that are associated with a conceptual taxonomy. But, as shown by the examples of section 2 above (and these to be given below), not all differences between synonyms and near-synonyms can be described in terms of such coarse restrictions.

However, many of the differences can be expressed in terms of various lexical features. For example, the difference between *glance* and *gaze* is the duration of the action. Textbooks on word usage (such as Room 1985 and Cornog 1992) and on translation (such as Vinay and Darbelnet 1958, Guillemin-Flescher 1981, and Astonigton 1983) have long recognized that lexical choice depends in part upon such features. Our claim is that it is possible to derive systematically a constrained (but not finite) set of such features that can be used to distinguish similar words, both across languages and within a single language.

4 A study of usage notes

Our claim arises from a study that we have made of dictionary usage notes. It is usually the explicit purpose of these notes to explain to the ordinary dictionary user what the differences are between groups of synonyms and near-synonyms. Figure 2 shows a typical example. By looking for regularities in the way that the notes explain the differences, we can determine what factors are important in lexical differentiation. The assumption is that although usage notes are given only for cases where the average dictionary user is likely to find difficulty, the terms in which the distinctions are made are nevertheless representative of lexical distinctions in general.

---

---

We used two English dictionaries in our study: an on-line copy of the *Oxford advanced learner’s dictionary* (OALD) (fourth edition, 1989) and a paper copy of the *Longman dictionary of contemporary English* (LDOCE) (second edition, 1987). In the first case, it was possible to extract all the usage notes automatically; in the second case, the notes were well-marked and easily recognized. (There were about 200 notes in the OALD, covering about 800 words, and approximately 400 notes in the LDOCE.)

We read through both sets of usage notes, studying the factors that were given to explain the differences between the words covered by each note. We observed that there were certain dimensions that were used quite frequently as denotative or connotative differentiae. Altogether, we noted 26 such dimensions for denotation and 12 for connotation (including a few that we added from the discussion of Vinay and Darbelnet (1958)). (We don’t, of course, claim this set to be complete or definitive.) Some of the dimensions are simple binary choices; others are continuous. Some examples are listed in figure 3. Each line of the table shows a dimension of differentiation (named, in most cases, for its endpoints), followed by example sentences in which two pleonys

---

---

usage information on synonyms and near-synonyms. Usage and translation guides such as Room 1985, Cornog 1992, and Hervey and Higgins 1992 also include this information, though they generally go beyond the requirements of the present study. Technical books can also be a source of information; for example, *Larousse gastronomique* (Montagné 1938/1961; Coutine 1984/1988) was a useful guide for us on the difference between *soupe* and *potage.* Unfortunately, sources can contradict each other, in which case one must make a judicious choice: for example, in the *soupe / potage* case just mentioned, the two editions of *Larousse gastronomique* contradicted each other and Hervey and Higgins (1992) in their exact differentiation of the terms.
Different Levels

In this section, we discuss our representation for a lexicon in which semantic and stylistic distinctions can be made between synonyms and pleononyms, both within and across languages. The central idea is that coarse denotational differentiation occurs at the language-independent conceptual level, and connotational and fine denotational differentiation occurs at the language-dependent level, in the lexical entries themselves. A key question is where exactly the best place is to draw the dividing line between the two levels.

5.1 The Conceptual Domain and the Lexical Domain

The starting point of our proposal is a familiar idea: a conventional KL-ONE-style taxonomic knowledge base serves to represent the basic semantic distinctions made by words in all the languages under consideration. (The implementation is in LOOM; see Stede 1993b.) The relations used in the KB derive from standard semantic case theory, and sentences are represented as usual: configurations of concepts and the relations that hold among them.

In simple, monolingual natural-language generation systems based on such representations, it is usual for concepts and words to be placed in direct correspondence: there is exactly one lexeme available to express each concept in the KB, thereby finessing any problem of lexical choice (see Stede 1993b for discussion). The KB is thus implicitly language-dependent and the finer grained it is, the greater the dependency—that is, the greater the number of changes that would have to be made to the conceptual taxonomy to replace the words with those of another language. Such an arrangement is probably not a good idea even in a monolingual system, and in multilingual applications, such as machine translation or multilingual generation, it is intolerable.

In trying to represent the meanings of the words of many languages simultaneously, such a conceptual hierarchy would not be language-independent but rather massively language-dependent—it would be the union of all language dependencies.8

But there has to be some place at which we slip from concepts to words. Our proposal here is that it should be earlier rather than later. Thus the conceptual hierarchy records, rather, the intersection of language dependencies and the fine tuning is then done at the lexical level for each separate language, even though the differentiae might ultimately be conceptual.

8This is exemplified by approaches like that of Emele et al (1992), who deliberately include concepts in the KB for every word in any of the target languages (and no other concepts!): “Each concept in this hierarchy has to have a lexical counterpart in at least one of the languages considered in the project . . . Conversely, each lexical unit of each language is related to a concept” (p. 66).
At the conceptual level, we represent the denotation of similar words in the KB by mapping them onto the same KB concept, but possibly with different thematic roles, restrictions, or distinguishing semantic traits. Therefore, we associate lexical items not to concepts only, but to entire configurations of a concept and various roles and fillers. (A similar proposal has also been made by Horacek (1990).) Furthermore, to achieve multilinguality, we apply the notion of near-synonymy across languages: pairs of equivalent or almost-equivalent words in different languages are seen as synonyms or near-synonyms, respectively.

For example, figure 4 shows the English and German words associated with the concept DIE—die, pass away, perish, kick the bucket, sterben, entschlafen, and abkutan—with the restrictions that pass away and entschlafen can apply only to people, and perish, kick the bucket, and abkutan can apply to people or animals but not plants (cf. Cruise 1986).

To establish the link between the concept, the EXPERIENCER role, and the appropriate filler (ANIMATE-BEING, ANIMAL, HUMAN) on the one hand, and the lexical item on the other, we create an instance of the concept, whose properties exactly reflect the conditions necessary for using the lexical item. These instances serve as the interface between the conceptual knowledge and the lexicon; they have roles pointing to the actual lexical entries for the languages used, wherein the connotational features and syntactic properties of the words are stored.

A more complicated situation arises when roles or role filler restrictions, as well as a concept, are part of the meaning of a word. This leads us to make a distinction between those parts of the KB that a word denotes and those parts that it covers; the latter might be only a subpart of the denotation. For example, the English verb heat and the German erhitzen both denote and cover just the concept APPLY-HEAT-TO. However, cook and kochen ‘prepare for eating by applying heat’ denote not only APPLY-HEAT-TO but also its PATIENT role and the selectional restriction of the role, FOOD;10 but they cover only the concept, not the role or its restriction. On the other hand, block and sieden, and a separate sense of kochen extend cook by adding the role HAS-GOAL-STATE with the filler BOILING, and both the role and the filler are included in what the word covers. Thus one may say Heat the milk until it is boiling or Boil the milk, but it is pleonastic to say Boil the milk until it is boiling. Similarly, fry and braten ‘cook over direct heat in hot oil or fat’ extend cook, but with the role INSTRUMENT and filler FAT, both of these being covered by these words.

We show our definitions for these words in figures 5 and 6. In figure 5, the coverage (not denotation) of each word is shown by the area of the dashed lines. Figure 6 shows the LOOM definitions, with the distinction between coverage and denotation. As before, the symbols in quotation marks are pointers to the complete lexical entries. Our earlier example of einschrauben, festschauben, and festschieben (see footnote 2) can be handled in a similar manner.

The effect of linking lexical items to concepts and roles is that we can represent more finely grained semantic distinctions than those made by the concepts only: similar lexical items all map onto the same, fairly general, semantic predicate, and the associated roles and fillers represent the smaller denotational differences.

To use this representation, we have developed a lexical option finder that traverses the proposition to be expressed and determines all lexical items that can denote some parts of the proposition. These items may vary in connotation and in precise denotation; later stages of the generation process will have to select from this pool the subset of items that is most appropriate to express the given message. (The lexical option finder is described in greater detail by Steede (1993b).)

5.3 The limitations of the conceptual level

Unfortunately, attaching words to taxonomized concepts has its limitations in dealing with linguistic nuance. The first problem that has to be dealt with is those cases in which a word applies to most but not all subordinates of some concept with which it is as-
(tell (:about
  heat_i APPLY-HEAT-T0
  (covering heat_d)
  (e-lexeme "heat")
  (g-lexeme "erhitzen")
)

(tell (:about
  cook_i APPLY-HEAT-T0
  (patient food_d)
  (covering cook_d)
  (e-lexeme "cook")
  (g-lexeme "kochen1")
)

(tell (:about
  boil_i APPLY-HEAT-T0
  (goal-state boiling_d)
  (:filled-by covering
    boil_d goal-state_d boiling_d)
  (e-lexeme "boil")
  (:filled-by g-lexeme "sieden" "kochen2")
)

(tell (:about
  fry_i APPLY-HEAT-T0
  (instrument fat_d)
  (:filled-by covering
    fry_d instrument_d fat_d)
  (e-lexeme "fry")
  (g-lexeme "braten")
)

Figure 6: LOOM instances for denotation and coverage of verbs of cooking.

Figure 5: Coverage of the conceptual hierarchy by different English and German verbs of cooking.

associated. For example, the German ausbessern applies to inanimate objects except for engines and machines (Schwarze 1979, p. 322). There are, generally, three ways of dealing with this kind of situation. First, one could introduce a new level into the concept hierarchy below INANIMATE-OBJECT and separate MACHINE from OTHER-INANIMATE-OBJECT. This step has an ad-hoc flavor to it; but the reluctance to taking it can be overcome if other words turn out to make the same distinction. If not, the specific idiosyncrasy can be dealt with either on the conceptual level by barring the general verb (here, *ausbessern*) from percolating downwards to one particular branch, (here, MACHINE), or—if the idiosyncrasy does not pertain to semantic traits—on the word level by stating a collocational constraint, thereby leaving the word-concept mapping unaffected.

The second problem is that, as we saw in section 2 with the example of forest and wood and their cognates in other languages, many of the semantic distinctions that we want to make do not lend themselves to easy taxonomic differentiation. We would have to include in our taxonomy under TRACT-OF-TREES such concepts as SMALLISH-TRACT-OF-TREES and BIGGER-TRACT-OF-TREES, NEAR-CIVILIZATION, and so on, which are not taxonomically well motivated. Worse, we would have to include language-specific concepts with no clear interrelationship; e.g., BOS-SIZED-TRACT-OF-TREES and HOLZ-SIZED-TRACT-OF-TREES.

Third, as we also saw in section 2, much lexical differentiation lies in emphasis rather than conceptual denotation; recall the examples of organize / arrange and enemy / foe.

Although these situations can be dealt with, they do highlight the fact that the strength of the conceptual approach is also an inherent weakness: the differences between plesionyms are represented as differences between concepts, and this is not always easy or natural.
It is this weakness of the pure taxonomic approach that leads us to the second component of our lexical representation: explicit differentiation of words, or, intuitively, formalized usage notes. Rather than trying to represent all lexical distinctions as conceptual distinctions, we may include in the lexical entry associated with a configuration of concepts lexical-choice rules that describe the distinctions between several words associated with the concept configuration, very much as dictionary usage notes do. Thus in figure 4, there would be only a single lexical entry for both perish and kick the bucket, whose formal usage note would describe the factors (in this case, the difference in formality and in attitude to the deceased) that are required to choose between the two words. Similarly, in figure 6, there would be only a single lexical entry for both sieden and kochen. And TRACT-OF-TREES would not need to be refined any further (at least, not for this purpose); instead, a usage note for each language would describe the relevant lexical distinctions. We are just beginning our development of this idea. This section describes the approach that we are taking.

We observe that dictionary usage notes have a characteristic structure:

- a description of the factors that distinguish each word in a set of synonyms or near-synonyms;
- an example of the use of each word in the set.

The descriptions of distinguishing factors follow a style or ‘language’ particular to the notes. The elements of the language include the denotative and connotative dimensions and features that we described above (see figure 3); an infinite (but constrained?) class of emphases, and a set of ‘operators’ such as most general, most usual, mostly used, not normally used, neutral word, strong, emphasizes, suggests, and usually associated with. Each example in a dictionary usage note is either a single ‘exemplar’ or several ‘best exemplars’ (cf Smith and Medin 1981, Smith 1989)—that is, one or more typical instances of uses of the word.

Our intent is to develop a formal, computationally usable representation of usage notes that mirrors this structure, and that approaches the full expressive power of dictionary usage notes. Thus we are designing separate representations of usage descriptions and exemplars, defining the semantics of the relationship between these representations, and, in tandem, developing a process that would use these representations as part of lexical choice in generation of target text (and later, we hope, in stylistic analysis of source text as well). This work will be founded on a formalized version of the language of usage notes; the usage descriptions will draw upon our catalogue of features and emphases, as well as concepts in the hierarchy, while the exemplars will be constructed from the concepts.

6 Conclusion

We have described our current, continuing research on representing nuances of meaning and style in language, and applying the representation in machine translation and text generation. The key to our approach is to discover, with the aid of dictionary usage notes, just how word senses can subtly differ, and to then use such features in a conceptually-based lexicon in which the finest-grained differentiation is made by formal usage notes.

Acknowledgements

For encouragement, inspiration, and clever ideas, we are indebted to Frank Toutai and Eduard Hovy. Some of our French data is from Anne Marie Miraglia, and Danish and Dutch from Henry Scholt; we are grateful to both of them. Our research is supported by grants from the Natural Sciences and Engineering Research Council of Canada and the Information Technology Research Centre. Some of the work described in this paper was carried out during the third author’s internship at the Research Institute for Applied Knowledge Processing (FAW), Uln, Germany, with the support of a travel grant from the International Computer Science Institute, Berkeley.

References


