Referência:
LEFFA, Vilson J. Word sense disambiguation in reading comprehension. In: MEYER, Françoise; BOLÍVAR, Adriana; FEBRES, Judith; SERRA, Marisela Bonett de. *English for Specific Purposes (ESP) in Latin America.* Mérida, Venezuela: Universidad de Los Andes 1997. p. 168-173

WORD SENSE DISAMBIGUATION IN READING COMPREHENSION

Vilson J. LEFFA, Catholic University of Pelotas, Brazil

Introduction

One of the basic assumptions in reading comprehension research is that the understanding of a text involves the cooperation of different knowledge sources. These knowledge sources reside in the reader's mind and have to be activated so that the text can be understood. Among these sources, one of the most important is the ability to recognize the meaning of the different lexical items that make up the text. This is a complex task in itself and one aspect that adds up to this complexity is the fact that words have many different meanings, out of which, usually, only one can be chosen. The instruments available to the reader to arrive at the one meaning supported by a given text, is the point addressed in this paper.

The following example, taken at random from *The Columbia Dictionary of Quotations*, can be used to demonstrate the complexity involved in assigning meaning to each of the lexical items.

> Consul. In American politics, a person who having failed to secure an office from the people is given one by the Administration on condition that he leave the country. (Ambrose Bierce (1842–1914), *The Devil's Dictionary*)

Looking up each word in The American Heritage Dictionary, and counting their meanings produces the following results (shown in brackets):

> Consul(3). In(22) American(6) politics(6), a(25) person(9) who(3) having (15) failed(12) to(21) secure(16) an(3) office(8) from(5) the(8) people(9) is (10) given(21) one(13) by(16) the(8) Administration(7) on(25) condition (12) that(15) he(2) leave(13) the(8) country(8).

The average meaning for each word is 11.34. If the meaning of each word depended on the meaning of every other word, we would have, in the small passage above, the astronomical sum of 11^{29} possibilities. But language does not work this way. Words do not multiply their meanings when in use, but have them restricted, typically to only one. The simple presence of other words, either to the right or to the left, has an inhibiting effect. No matter

how many meanings a word has in the dictionary, it can carry only one when in the company of other words in a text. The pressure from the text is so great that the word may even loose its individuality, sometimes to the point of being divested of all its original dictionary meanings and imposed a new one.

A problem faced by readers when dealing with a text in a foreign language is that the different meanings a word presents in the new language are not symmetrical to the ones presented in the first language. This happens with practically every word. The apparently unambiguous *he*, in the passage above, when translated into Portuguese, has a different rendering in each of the following sentences (translation in brackets):

> The cat is a he (macho).
> He who seeks equity must do equity (aquele).
> It was he who pushed the controversial points (ele).

What kind of knowledge sources does the L2 reader use to deal with this problem? The traditional approach is to emphasize the role of world knowledge (Minsky, 1975; Shank & Abelson, 1977; Rumelhart, 1981). When an ambiguous word is found, the reader solves the ambiguity by activating the adequate schema. In the passage above, for example, the ambiguous word *office* would be interpreted as a public position, not room or building, because the election or political schema is invoked.

The problem is that schemas, to the extent that they represent larger chunks of knowledge, are not refined enough to solve many of the ambiguities that are encountered when two languages are involved. The verb *fail* in the sentences below can refer to the same all-including negociations schema, but in each sentence it has a different meaning.

> The delegates failed
> The delegates failed to reach an agreement.
> The delegates cleverly failed to show the statistics.
> The delegates failed their own people.

The negociations schema, although adequate, does not seem to be specific enough to discriminate between all the meanings. Some other resource may have to be used. The hypothesis advanced here is that resorting to world knowledge is not necessary to resolve lexical ambiguity in a normal, authentic text. Linguistic knowledge of the possible meanings of a word, combined with the syntactic and semantic restrictions, probably produces faster and more economical results. It is even possible that by the time the reader activates a given schema or frame, the ambiguity is already resolved.

In a previous study (Leffa, 1996), we showed that L2 readers were able to solve lexical ambiguities by resorting exclusively to syntactic and semantic restrictions. These restrictions would give the word “condition”, for example, one meaning when preceded by “on” (“on condition”) and a

different meaning when preceded by “heart” (“heart condition”). Since the study, however, used human subjects it was not possible to prevent them from using world knowledge, even if the textual constraints proved that it was not necessary.

This study was designed to use a computer program instead of human subjects. The methodological advantage is that restrictions based on syntactic and semantics relations can be totally isolated from world knowledge, as based on schema activation. Another advantage is that a larger, more reliable measuring instrument can be used, involving thousands of items — something which is unfeasible with human subjects.

Methodology

The methodology used to collect the data involved three steps: (1) selecting a list of ambiguous words; (2) compiling examples of use for each word from a corpus; (3) processing the examples in a computer system, using syntactic and semantic constraints.

The first step was making up a list of words in English that would produce different translations in Portuguese. The first criterion for selecting these words was that they belonged to the same part of speech: ambiguities between “answer” as a noun and “answer” as a verb, for example, were discarded. Another criterion was that the ambiguous words should not depend too much on the immediate context to have the ambiguity solved such as prepositions, which are usually resolved at the syntactic level (“I depend on you”). The reason for trying to select words that depended on a larger context to be disambiguated was the assumption that it would result in a more reliable test of the hipothesis. The part of speech that seemed to depend less on the immediate context, all other things being equal, was the noun. The final shortlist included the following target nouns: arm, bank, bar, bill, board, chip, coat, coach, corner, driver, gum, letter, nail, page, plane, record, room, table, time, wall.

The second step was compiling examples of use for each word. The source for these examples was a corpus of 20,000,000 words of expository text. Occurrences of each word were recorded using the *Oxford Microconcord* (Scott, 1992). Since many of the ambiguous words belonged to different parts of speech, any part other than the noun was discarded. The occurrences were finally reduced to 200 examples for each word, using a random selection procedure, which resulted in a total of 4,000 examples. Each example was 140 characters long, producing segments of text with about 20 words each.

The third step was disambiguating the target nouns in the examples. A computer system that is being developed for machine translation was used for this purpose. This system allows for disambiguating rules to be introduced at a certain stage of the translation process. Figure 1 shows an example of such a rule to disambiguate between the two meanings of “left” in the sample sentences. As can be seen, the rules do not assume any world

knowledge of the type described in schema theory.
The 4,000 segments of text with the target words were fed into the system and processed. The output was a tentative Portuguese translation of these English segments, at the lexical level. The morphological attributes and syntactical rules of the Portuguese language were not included here.

Sentence 1:	When he left the house he was ready.
Sentence 2:	When he left the house was ready.
Ambiguous word:	left
Disambiguating rule:	if “left” is a verb and the following NP is the object of the verb, then translate “left” as “deixar”; if left is a verb and the following NP is the subject of another verb, then translate “left” as “partir”.

Figure 1: Example of a disambiguating rule.

Analysis
A first look at the output (Figure 2) shows that, in terms of translation, a lot of garbage was produced. The original English text, segmented arbitrarily in chunks of 140 characters, produced incomplete sentences and even incomplete words. As the segments were all put together, the program treated them as parts of the same text, connecting parts that should not be conneted. Besides the 20 target words examined here, there were also many other ambiguities in the textual segments for which the program was not yet prepared to cope with. Although all these difficulties put together place an unfair demand on the system itself, it is also believed that in terms of this investigation they contribute to a more robust testing of the hypothesis.
Table 1 shows the results for the 4,000 segments with the 20 ambiguous words. Resolution rate varied from 85% to 100%, providing an average of 96.55% of correct disambiguations, with a low standard deviation, which means that results tended to be similar. More variation was noticed in terms of distance between the ambiguous word and the term that disambiguates it, which we will refer to as *collocate*. Examples of collocates are the words *hand* and *munitions*, in Figure 1, used by the system to disambiguate *arm*.

...kness, spasticity, and atrophy, usually starting in the hands and *arms* and then spreading to other parts of the body. Difficulty with speak...	...kness, spasticidade, e atrofia, geralmente começar no mão e *braço* e então espalhar para outro parte do corpo. Dificuldade em falar...

...hat aid took the form of the government's handing over munitions, *arms*, and clothing to the playwright Caron de BEAUMARCHAIS and his fake "H...	...chapéu auxílio tomar o forma do governo estar entregar munição, *arma*, e vestuário para o dramaturgo Caron de BEAUMARCHAIS e seu fake "h...

Figure 2 — Examples of disambiguated target words (in italics)

A given word may belong to a phrase, which was arbitrarily separated into corpus phrases, based solely on frequency of occurrences in the corpus, and dictionary phrases, usually based on meaning. Examples of dictionary phrases with *arm* include: *arm and leg, arm of the law, arm's length,* etc. Examples of corpus phrases are: *take up arms, bear arms, left arm, small arms.* Corpus phrases may or may not coincide with dictionary phrases.
As expected, there is a high correlation (r = .92) between resolution and distance from the collocate. As distance increases, intervening factors can affect the results. Distance and direction of the collocate (right or left) are important factors in deciding how to disambiguate a word, as can be seen in Table 2, which displays the collocates for arm (COBUILD, 1995). Proficient users of the English language would intuitively expect *arms control* to mean *gun control* and *arms around* as something related to *embracing*.
The average of 96.55 of correct disambiguations reflects the applications of rules as they were introduced into the system. Some of these rules can be improved and eventually produce better results — as can be seen in the following sentence:

Table 1 — General results

Target words	Resolution %	Distance in words	Phrases in corpus	Phrases in dictionary
arm	100	1.5	4	8
bank	100	2.0	8	13
bar	89	6.0	9	11
bill	99	2.1	8	14
board	95	2.5	12	15
chip	98	3.5	11	2
coat	99	3.0	10	6
coach	100	2.0	7	2
corner	85	8.0	11	6
driver	95	4.0	6	3
gum	100	2.1	10	7
letter	94	4.1	2	16
nail	100	2.2	5	9
page	99	2.0	3	5
plane	99	2.1	11	6
record	98	2.9	9	18

room	94	6.0	7	10
table	97	3.7	7	17
time	98	2.0	7	39
wall	92	7.0	5	9
mean	96.55	3.44		
SD	4.09	1.89		

Table 2 — Most frequent collocates for arm

Collocates	Frequency	Collocates	Frequency
control	1485	embargo	607
around	719	down	518
legs	709	nuclear	516
sales	622	treaty	426

> The music is set in duple metre (2 beats to the bar) and is based on about 50 standard calypso melodies.

In the sentence above, *bar* was incorrectly interpreted by the system as a room where drinks are served and music can be played, instead of the measure used in music to divide a staff. The problem was that the collocate *music* was not specific enough to discriminate between these two meanings of bar. There are, however, other more restricting collocates in the sentence — which could have been fed into the system to solve the problem such as *duple metre* and even *beats*, although an ambiguous word itself.

Conclusion

Words, before they are used in a text, are just a set of possibilities, pointing imprecisely to a bank of concepts we have stored in dictionaries or in our minds. In terms of dictionaries, the number of meanings assigned to every word, as they are used in a current text, is around 11 meanings per word. In terms of what we have stored in our minds the number is probably much higher, including hundred or maybe thousands of recessive meanings, meanings that are hidden behind the dominant one and that come to life when certain conditions are met in a text. Once, however, a recessive meaning becomes dominant all the others become recessive, discounting for the rare cases when double meaning is intentionally used. The main finding in this investigation is that this drastic reduction to one meaning is due to syntactic and semantic restrictions imposed by the neighboring words. It is argued that this is advantageous to the process of reading, making it more efficient.

Of course we can always build examples in which a given word may be disambiguated only by the activation of a given schema. The data analyzed here, however, suggest that in cases where both schemas and textual

restrictons can be used, the application of syntactic and semantic restrictions is more precise and economical.

It can also be argued that once a schema is activated it guides disambiguation, sometimes to the point of predicting what is coming next. If somebody is reading a text about hand care and meets a sentence that starts with the words “when you cut your ...”, the meaning of *nail*, as part of the finger, is probably activated even before the word is read.

The problem here, it seems, is to decide which comes first, schema activation or data from the text. It is true that in some cases schemas may be previously induced, such as in a classroom situation where the teacher prepares the student for the reading of a text. Most often, however, it seems that schemas are activated as data are processed from the text.

Another argument that could be used to favor schemas against textual constraints is that schemas are more powerful to help the reader guess the meaning of unknown meanings for ambigous words. The reader, for example, may be familiar with the meaning of pen as a writing instrument and may have problems when he or she meets the word used in the sense of an enclosure for animals.

Again, it seems that the reader does not need the broader context provided by a schema to guess the new meaning of the word. A list of short examples would probably be more helpful.

The use of syntactic and semantic constraints, as compared to world knowledge, occur automatically, below the level of consciousness. The sequential and more time-consuming strategies are replaced by faster, automatic processing, where activities are performed in parallel. The result, whenever we move a subprocess to these lower, more automatic levels, is a general gain in reading efficiency.

REFERENCES

COBUILD, *English Collocations on CD-ROM*. (1995) London: Harper Collins.

MINSKY, M. (1975) A framework for representing knowledge. In: WINSTON, P. (Ed.) *The psychology of computer vision*. New York: McGraw-Hill, p. 211-277.

RASKIN, V. (1987) Linguistic heuristics of humour: a script-based semantic approach. *International Journal of the sociology of Language*, v. 65, p. 11-25.

RUMELHART, D. E. (1981) Schemata: The building blocks of cognition. In: Guthrie, J. T. (Ed.) *Comprehension and teaching: Research reviews*. New Haven: Internation Reading Association.

SCHANK, R. C. & ABELSON, R. P. (1977) *Scripts, plans, goals, and understanding*. Hillsdale, N. J.: Lawrence Erlbaum.

SCOTT, M. & JOHNS, T. (1992) *Microconcord.* Oxford: University

Press. (Oxford English Software).

WEINER, J. & DE PALMA, P. (1993) Some pragmatic features of lexical ambiguity and simple riddles. *Language & communication*, v. 13, n. 3, p. 183-193