CONCEPTUAL DISTANCE AND AUTOMATIC SPELLING CORRECTION

E. Agirre, X. Arregi, X. Artola, A. Díaz de Ilarraza, K. Sarasola

Informatika Fakultatea, p.k. 649. 20080 DONOSTIA
(Basque Country - Spain)
e-mail: jibagbee@si.ehu.es
tel: 34 43 218000

ABSTRACT. Text from different sources usually arrives under imperfect conditions. When an anomalous word is detected automatic word recognisers produce a list of candidates from which only one is correct. A variety of techniques have been devised to discriminate among the possible correction candidates. The project we are involved in tries to exploit linguistic knowledge in Spelling Correction. A preliminary investigation shows syntactic discrimination not to be enough. The gap could be covered by semantic techniques like conceptual distance. Basically, we define conceptual distance between two concepts as the shortest path length in the hierarchies of the lexical knowledge base of IDHS (Intelligent Dictionary Help System). We consider that a correction proposal that is closer to the surrounding words in the sentence is more plausible enabling us to produce a ranking of the proposals. It is our belief that conceptual distance can be also applied to other word recognition areas, such as handwriting recognition or optical character recognition, where a single proposal would also be desirable.

1 INTRODUCTION

Text from different sources usually arrives under imperfect conditions. The medium of transmission conditions the type of automatic word recognition to be used: Optical Character Recognition, Speech Processing or Spelling Correction. When an anomalous input is encountered these recognisers produce a list of candidates from which only one is correct. There are a number of applications e.g. Text-to-Speech Synthesis, that in order to rule out human intervention need automatic correction, that is, the first choice of the correct proposal among the correction candidates.

The task of choosing the appropriate correction proposal is not an easy one. We have to draw knowledge from several sources, as one technique alone would not suffice. In this direction [Kukich, 92] points out, for spelling correction and considering isolated words only, that automatic correction performed by humans scored from %65 to %82. These figures could represent an upper bound for automatic techniques that do not take context into account. To leave %35-%18 of the detected errors uncorrected would be unsatisfactory for the applications mentioned earlier. In order to increase the performance and get an acceptable correction rate, some sort of context modelling, linguistic or other, would be needed.

The project we are involved in tries to exploit linguistic knowledge for automatic spelling correction. This paper focuses on the contribution of lexical-semantic techniques in general, and conceptual distance in particular. Some other work is being carried on the syntactic side.

The idea of conceptual distance captures the intuition that some words are more related or closer than others. We consider that a correction proposal that is closer to the surrounding words in the sentence is more plausible. Thus we can produce a ranking of the proposals.

Basically, we define conceptual distance between two word senses as the shortest path length in the hierarchies of the Dictionary Knowledge Base of IDHS (Intelligent Dictionary Help System [Artola, 93; Agirre et al., 94]), following the ideas of [Rada et al., 87]. The knowledge base of IDHS is a semantic network of frames where each frame represents a word sense from a dictionary. Arcs between frames represent lexical-semantic relations derived from the definitions in a machine readable dictionary.

Next section shows some experimental results that indicate the need of more linguistic knowledge beyond syntax in spelling error correction, followed by an overview of IDHS. After that, two prospective semantic techniques are introduced, from which conceptual distance is explored in depth in the next section. Finally some conclusions are presented.

Originally, the target language was Basque, but later developments in IDHS made us switch to French. For this reason the preliminary collection of data was done for Basque, while the implementation is being run on French texts. The examples in sections 2 and 4 are in Basque, while those in section 5 are in French.

2 ON THE NEED OF SEMANTIC DISCRIMINATION

In order to have some hard data on the convenience and prospective performance of the semantic contribution to automatic error correction, the analysis of a small corpus was performed. The error detection and the list of proposals have been taken from the spelling checker/corrector XUXEN [Aduriz et al, 1993; Agirre et al., 1992]. The texts come from 48 Basque language learners, giving a total of 8290 words. XUXEN generated proposals for 305 spelling errors, producing multiple proposals 182 times (60%).

The syntactic analysis of the texts, as well as the syntactic discrimination of the proposals, was performed by a person simulating an automatic full-fledged and robust parser. The proposals which would lead to grammatical errors where thus removed from the proposal lists. The semantic discrimination was applied only after the syntactic phase was completed.

The results hold that syntax alone could select one single proposal 70% of the cases. This result might be too optimistic, considering that the syntactic analyser was supposed to be complete and robust.

The semantic information faced the cases where syntax could not do the job. Applying by hand the semantic techniques explained below, it managed to solve 63% of the misspellings. It might be that this experiment favoured syntax, leaving semantics the tough cases. Anyway, the performance of both is similar, and the experiment indicates that their combination is desirable in order to get better results, up to 90% in this particular experiment. These results are tentative, awaiting confirmation of implemented systems with realistic syntactic and semantic coverage.

XUXEN: 305 errors with proposals		
1 prop.	123	40.3%
n prop.	182	59.7%
syntactic discrimination on 182 errors		
success	128	70.3%
fail	54	29.7%
semantic discrimination on 54 errors		
success	34	62.9%
2/3	11	20.3%
fail	9	16.8%

3 IDHS

IDHS (Intelligent Dictionary Help System) provides the base for semantic correction. It provides both a representation language suited to explore the techniques presented in the following section, and also the semantic knowledge itself.

IDHS was conceived as a monolingual (explanatory) dictionary system for human use [Artola & Evrard, 92; Artola, 93]. The system provides various access possibilities to the data, allowing to deduce implicit knowledge from the explicit dictionary information. The system has been implemented on a symbolic architecture machine using KEE knowledge engineering environment.

The starting point of IDHS is a Dictionary Database (DDB) built from an ordinary French dictionary. Meaning definitions have been analysed using linguistic information from the DDB itself and interpreted to be structured as a Dictionary Knowledge Base (DKB). As a result of the parsing different lexical-semantic relations between word senses are established by means of semantic rules (attached to the patterns); rules are used for the initial construction of the DKB.

The interconceptual lexical-semantic relations detected from the analysis of the source dictionary are classified into paradigmatic and syntagmatic. Among the paradigmatic relations, the following have been found: synonymy and antonymy, taxonomic relations as hypernymy/ hyponymy —obtained from definitions of type "genus et differentia"— and taxonymy itself (expressed by means of specific relators such as sort-of and kind-of), meronymy, and others. Whereas among the syntagmatic relations we can find case relations (e.g. agent, object, goal, etc.), relations derived from the specific lexicographic metalanguage (e.g. quality-of, act-of, property), and others.

The knowledge representation scheme chosen for the DKB of IDHS is composed of three elements, each of them structured as a different knowledge base. One of this components, KB-THESAURUS, implements the dictionary as a semantic network of frames, where each frame represents a one-word concept (word sense) or a phrasal concept. Phrasal concepts represent phrase structures associated to the occurrence of concepts in meaning definitions. Frames are interrelated by slots representing lexical-semantic relations. Other slots contain phrasal, meta-linguistic, and general information.

In the following section we tackle spelling correction from the point of view of semantics and IDHS.

4 SEMANTIC DISCRIMINATION

As we already mentioned, this work focuses primarily on the contribution of semantics, and more precisely in the use of lexical-semantic information. We have considered the use of the following:

Selectional Restrictions

Selectional restrictions indicate semantic constraints that the arguments of verbs, adjectives or nouns have to fulfil. For example:

```
jan       => verb[agent: animate,
                  object: edible]
ilegorri  => adj.[argument: person]
anaia     => noun[argument: person]
```

These can be read as 'the verb `jan` (eat) takes as agent an animate entity and as object and edible entity', 'the argument of `ilegorri` (blonde) has to be a person', etc.

The contribution of selectional restrictions will be illustrated by the following example from the Basque corpus. Had someone typed `lehio` in Basque we would get the proposals below[1]:

```
lehio: lehia, lesio, leiho
```

If the misspelling occurs in the following sentence, and assuming a sample selectional restriction for `apurtu` (to break),

```
"lehio bat apurtu dut"[2]
apurtu =>  [agent: animal,
            object: physical-object]
```

we would be able to discard competition and injury, and select the only proposal that fulfils the restriction of being a physical object, `leiho` (window).

Conceptual Distance

The idea of conceptual distance tries to capture the intuition that some words are closer or more related than others. Therefore we can consider devising a metric that would give results similar to the following[3]:

```
dist(itsasontzi,kapitain) = "short"
dist(itsasontzi,teklatu) = "long"
```

The idea is that we prefer proposals that are related or conceptually close to the other words in the sentence, rather than unrelated or distant proposals. This approach has multiple variants, depending on whether we take all the words in the sentence, or we only take the measurements with some relevant words in the sentence.

Let us consider the following example[4]:

```
uzaina: zaina, usaina, uhaina
"ukenduaren uzainak erlea aldendu zuen"
```

We can compare the distance of the proposals with the other words in the

[1] The proposals mean respectively competition, injury, window.

[2] Meaning *I broke a <lehio>*. All the basque examples and proposals in the paper are taken from a small corpus and the correction proposals are all from Xuxen

[3] The words mean respectively *ship, captain, keyboard.*

[4] The proposals mean, respectively, *vein, smell, wave.* The sentence means *the <uzaina> of the ointment kept away the bee.*

sentence. The result would be that usaina (smell) holds the minimum total distance, and therefore would be preferred as the correct proposal. This technique will be further explained below.

5 CONCEPTUAL DISTANCE AND SPELLING CORRECTION

Mainstream approaches to conceptual distance rely on structured inheritance nets or similar kinds of knowledge bases. For instance, [Rada et al., 89] defines conceptual distance in terms of the length of the shortest path of IS-A links between the word senses of the Mesh semantic net. Besides applying distance in a medical bibliographic retrieval system, they also try to use it as a tool for merging semantic nets.

In a similar approach, [Sussna, 93] assigns a weight to each link in the Wordnet semantic network and calculates the distance between two word senses as the total weight of the path with minimum weight. The weights try to capture additional data, e.g. tfor the same path length, word senses lower in the hierarchy seem to be conceptually closer.

These two approaches take into consideration that words have multiple senses. In fact [Sussna, 93] devises his measure with the purpose of sense-disambiguating a text for indexing and text retrieval.

The knowledge representation of IDHS provides support for the experimentation of several distance measures, allowing us to select the most suitable for proposal discrimination. Previous works on conceptual distance rely mainly on hierarchical relations (hypernymy, taxonymy, meronymy), but distance measures could also profit from the other semantic relations in IDHS. [Rada et al., 89] point out that the proliferation of semantic relations makes distance unreliable. Such systems (e.g. [Collins et Loftus, 75]) have to provide a complex weighting mechanism to balance the heterogeneous nature of the relations. In order to avoid that, it would be desirable to use certain semantic relation only when appropriate, that is, when it makes sense in the given context. This idea will be developed below, while considering the issues related to the application of conceptual distance to correction.

Path-Finding Algorithms

In the heart of the distance algorithm there is a path-finding algorithm. Given two word senses in IDHS, the algorithm would find the shortest path(s) of lexical-semantic links between both. In order to be able to test different correction strategies the following algorithms have been implemented:

h-path(n1,n2): finds the path following hierarchical links only: hypernym, part-of, component-of, element-of, sort-of and their respective inverse relations.

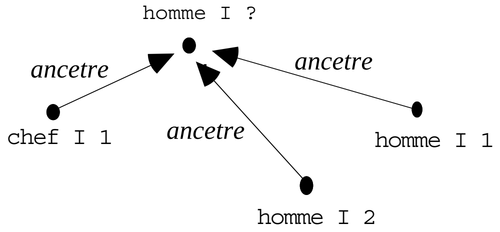
s-path(n1,n2,r1,...,rn): finds a path that has to contain at least one non-hierarchical (semantic) link from the set {r1,...,rn}, alongside the previously mentioned hierarchical links.

s*-path(n1,n2): finds a path that may contain any non-hierarchical (semantic) relation, alongside the hierarchical links.

The first algorithm, h-path, constraints the search to hierarchical relations only. It is considered the most reliable for conceptual distance, but it imposes several limitations. The two word senses need to be in the same hierarchy, which implies that h-path will never find a path across different parts of speech. For the same reason, it needs very comprehensive hierarchies, which are difficult to create or acquire. Other semantic relations could alleviate this, relating concepts across hierarchies.

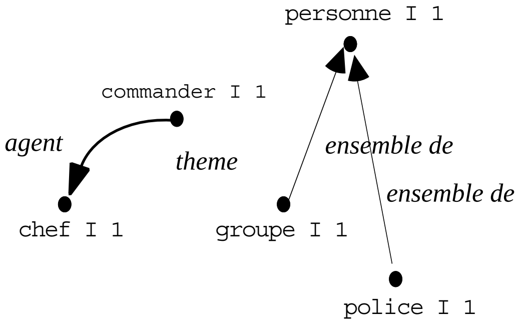
The use of unconstrained semantic relations as in s*-path, though, can produce nonsense paths that have to be neutralised when calculating the actual distance figures. It also has heavy efficiency burdens, which can be reduced constraining the set of acceptable relations. If the set of relations is constrained according to semantic criteria, the paths will be semantically coherent. The set of acceptable relations for a certain pair of word senses could be deduced from context, or in some cases, from the part of speech of the word senses. For instance, IDHS admits two relations for a noun that have an adjective as value: property and quality-of. In that case s-path will return a path that relates both noun and adjective via property, quality-of and the hierarchical relations.

Some examples of the algorithms follow:



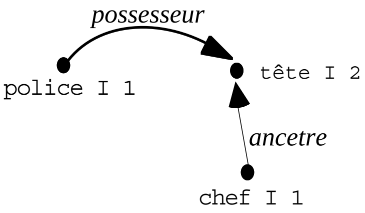
h-path(chefI1, hommeI1) =
chefI1 *ancetre* hommeI? *descendant* hommeI1

The path found by `h-path` between the first word sense of *boss* and the first sense of *man* means: *bossI1 is an ancestor*[5] *of manI?* (a non-disambiguated sense that includes all other senses of man), *which has as descendant manI1.*



s*-path(chefI1, policeI1) =
chefI1 *agent+inv* commanderI1 *theme* groupeI1 *ensemble de* personneI1 *element de* policeI1

The path found by `s*-path` between the first word sense of *boss* and the first sense of *police* means: *bossI1 is an agent of to-commandI1 which has as object groupI1, which is a set-of personI1 which is an element-of policeI1.*



s-path(chefI1,policeI1,possesseur+inv) =
chefI1 *ancetre* têteI2 *poss.+inv* policeI1

The path found by `s*-path` between the first word sense of *boss* and the first sense of *police* means: *bossI1 is a descendant of headI2 (as in head of department), which is "owned" by policeI1.*

The general search of a path between two nodes has exponential complexity, in the order of $O(c^n)$, where `c` is the average of the number of links per word sense, and `n` is the length of the path. In order to keep it under control, the length of the path has to be limited beforehand. This limit can be interpreted as the point after which we consider the two nodes to be unrelated or "very" far. Accordingly, this limit should be "tuned" having in consideration both efficiency and conceptual suitability.

The complexity of the three algorithms grows from the first to the last. While `h-path` deals with five hierarchical relations ($c \leq 5$) and `s-path` is devised to also take into account a small set of relations of the same kind (one to four extra relations, $c \leq 9$), `s*-path` has to provide for the whole set of relations (ranging from 10 to 40 depending on the part of speech of the word sense).

Conceptual Distance

The path(s) between two word senses is(are) the base for conceptual distance. But other facts have to be also considered. The empirical results of [Sussna, 93] show that, as already mentioned at the beginning of this section, the length of the path and the specificity of the word senses in the path (measured by the depth in the hierarchy) are the important parameters that affect the distance measure he proposes. The second parameter tries to capture the fact that specific word senses are considered closer than more general ones.

Our conceptual distance reflects those parameters in the following formula:

$$\mathbf{distance}(\boldsymbol{ws_1}, \boldsymbol{ws_n}) = \sum_{i=1}^{n} \mathbf{1/depth}(\boldsymbol{ws_i})$$

where $< \boldsymbol{ws_1} \ldots \boldsymbol{ws_i} \ldots \boldsymbol{ws_n} >$ is the path from $\boldsymbol{ws_1}$ to $\boldsymbol{ws_n}$, and $\mathbf{depth}(\boldsymbol{ws_i})$ is the depth of $\boldsymbol{ws_i}$ in the taxonomy.

Other parameters that could help tuning the measure have not been considered yet. One parameter, for example, could involve giving different weights to each relation, in a way similar to the "criteriality tags" used by [Quillian, 68]. The inclusion of these parameters in the above formula depends greatly on empirical results, which have not yet been gathered.

Correction

As mentioned in section 4, we perform correction choosing the proposal that is more

[5] Ancestor includes the concepts in the transitive closure of hypernymy. Descendant includes the concepts in the transitive closure of hyponymy.

related or conceptually closer to the other words in the sentence, and leaving aside unrelated or distant proposals. The relatedness of a given proposal with the surrounding sentence can be measured using a variety of strategies.

g-correction (generalised). Distance as defined above is measured between word senses. Consequently all the senses in the dictionary for the words in the sentence and the proposals have to be considered. This means that inappropriate senses could bias the corrector to choose an incorrect proposal. In order to rule out, or at least try to neutralise, these spurious readings, and at the same time choose the correct proposal, the following technique can be used: the preferred senses and proposals will be the ones that give minimal pairwise conceptual distance.

Thus, if we have a sentence of length N $<w_1, w_2, ...w_n>$ with M spelling errors $\{e_1=w_i...e_m=w_j\}$, and a list of proposals for each error $P(e_i) = <p_{i1},...p_{iL}>$, we need to consider the senses of all non-error words and the proposals. For each possible combination of senses (mixing both non-error words and proposals), the winning combination will be the one with the minimal total of pairwise distances. This winning combination will give both the preferred proposals and word senses.

In figure 1, it can easily be seen that for long sentences with highly ambiguous words and many correction proposals, the number of combinations and pairwise distance computations grows enormously.

c-correction (constrained). If we want to limit both the number of combinations and the pairwise distance computations, we can focus on doing proposal discrimination only. We are not trying to sense-disambiguate now, and will thus consider of equal value incorrect word senses and appropriate ones.

For each proposal we will only compute the distances of its corresponding word senses with each word sense of the non-error words in the sentence (cf. fig. 2). The proposal that gets the minimum total distance wins.

```
Sentence:   le cheé de la police reunit vingt hommes sur la place du village.
Error:  cheé                      Proposals:   chef cher chez chié chieé chéri chic

Word Senses in IDHS:
   Sentence:  police I 1, police I 2,
              reunir I 1, reunir I 2, reunir I 3, reunir I 4, reunir I 5
              homme I 1, homme I 2, homme I 3, homme I 4, homme I ?
              place I 1, place I 2, place I 3, place I 4, place I 5, place I ?
              village I 1
   Proposals: chef I 1, cher I 1, cher I 2, chéri I 1, chic I 1

Combinations:
   C1)        police I 1, reunir I 1, homme I 1, place I 1, village I 1, chef I 1
   C2)        police I 2, reunir I 1, homme I 1, place I 1, village I 1, chef I 1
       ...
     Number of combinations:  2x5x5x6x1x5 = 1.500

Distance on C1:
       dist(police I 1, reunir I 1) ... dist(police I 1, chef I 1)      n=5
       dist(reunir I 1, place I 1)  ... dist(reunir I 1, chef I 1)      n=4
       ...
       dist(village I 1, chef I 1)                                      n=1

     Number of distance calls:
              [total]   1500 x (5+4+3+2+1) = 1500 x 15 = 22.500
              [distinct pairs]                            239
```

fig. 1. Combinations in **g-correction**.[6]

[6] The sentence means "the *cheé* of the police gathered twenty men in the square of the village". The proposals for cheé are: boss, expensive, ´home of´, dear and stylishness..

```
Combinations:
   chef I 1  police I 1, police I 2,
             reunir I 1, reunir I 2, reunir I 3, reunir I 4, reunir I 5
             homme I 1, homme I 2, homme I 3, homme I 4, homme I ?
             place I 1, place I 2, place I 3, place I 4, place I 5, place I ?
             village I 1
   ...
   chic I 1  police I 1, police I 2,
             reunir I 1, reunir I 2, reunir I 3, reunir I 4, reunir I 5
             homme I 1, homme I 2, homme I 3, homme I 4, homme I ?
             place I 1, place I 2, place I 3, place I 4, place I 5, place I ?
             village I 1

      Number of combinations: 5

Distance:
   C1)       dist(chef I 1, police I 1) ... dist(chef I 1, village I 1)
             ...
             dist(chic I 1, police I 1) ... dist(chic I 1, village I 1)

      Number of distance calls:
         [total]     5x(2+5+5+6+1)= 95
```

fig. 2. Combinations in **c-correction**.

Although the wrong word sense may contribute to credit incorrect proposals, the greater number of related true senses will add up and eventually the correct proposals will be chosen.

s-correction ("*semantic*"). We have already introduced two path-finding algorithms (s-path and s*-path) that traverse non-hierarchical semantic relations. The semantic clues in the sentence can be used to inform s-path about the relations that can be expected in the path between the two word senses. Figure 3 illustrates a simplified example of the semantic relations in the sentence from figure 1. The preposition *de* can be interpreted as meaning owner, location etc. For the example below, calling s-path with the corresponding word senses will find a path. We already saw an example when examining path-finding.

This kind of semantic interpretation does not require as heavy a linguistic machinery as it might seem. Triples like those of the example are readily obtained by semantic information extraction systems from corpora [Velardi et al., 91].

6 CONCLUSIONS AND FURTHER WORK

We have outlined the application of a specific semantic technique, conceptual distance, in automatic spelling correction.

```
Semantic relations:

   from the verb:
   (reunit agent cheé)
   ...
   from the preposition de:
   (cheé possesseur+inv police)
   (cheé location police)
   ...
Combinations & Distance:
   reunir I 1 chef I 1...chic I 1
   ...
   reunir I 5 chef I 1...chic I 1

   chef I 1...chic I 1 police I 1
   chef I 1...chic I 1 police I 2

   ...

   Number of combinations:   5+2+2=9
   Number of dist. calls:       9x5=45
```

fig. 3. Combinations in **s-correction**.

In previous implementations of conceptual distance, only h-path style algorithms have been used. These algorithms need comprehensive hierarchies, which are difficult to construct. Other semantic relations. i.e. non-hierarchical relations, can serve to relate word senses even if they do not share the same hierarchy, and specially in the case of two word senses from different grammatical categories. These extra semantic relations could be exploited by conceptual distance using s*-path and s-path. Selectional restrictions are also an alternative in this kind of situations.

s*-path has coherence and efficiency problems which are alleviated in s-path. But in order to use s-path properly, semantic information from the context of the error has to be obtained. This semantic analysis and the tuning of the specific relations needed in a certain context are the work we are focusing on now.

In a further step, we are also planning to develop a more efficient application-oriented representation of the semantic knowledge. For that purpose, we will try to identify and map the relevant subset of the representation of IDHS.

Other important issue is the application of the different correction strategies to real data, where their performance should be effectively contrasted. In this sense, IDHS, because of the rich variety of semantic relations extracted from the dictionary, is very well suited as a platform for extensive testing of the issues above.

It is our believe that the correction techniques explored in this paper, although originally designed for spelling correction, are not dependent of the error source. As long as they are applied on linguistic input they could be used in other word recognition areas where automatic correction, i.e. single correction proposals, would be desirable.

ACKNOWLEDGEMENTS

Eneko Agirre was enjoying a grant from the Basque Government during the present work.

REFERENCES

Aduriz, I., Agirre, E., Alegria, I., Arregi, X., Arriola, J.M., Artola, X., Díaz de Ilarraza, A., Ezeiza, N., Maritxalar, M., Sarasola, K. and Urkia, M. A Morphological Analysis Based Method for Spelling Correction, in *Proceedings of the E.A.C.L.*, Utrecht, The Netherlands. 1.993

Agirre, E., Alegria, I., Arregi, X., Artola, X., Díaz de Ilarraza, A., Maritxalar, M., Sarasola, K. and Urkia, M. XUXEN: A Spelling Checker/Corrector for Basque Based on Two-Level Morphology in *Proceedings of the third conference on Applied Natural Language Processing*, Trento, Italy. 1992.

Arregi X., Artola X., Díaz de Ilarraza A., Evrard F., Sarasola K.. Aproximación funcional a DIAC: Diccionario inteligente de ayuda a la comprensión, *Proc. SEPLN*, 11, 127-138. 1991.

Artola, X. HIZTSUA: Hiztegi-sistema urgazle adimendunaren sorkuntza eta eraikuntza / Conception d'un système intelligent d'aide dictionnariale (SIAD). PhD thesis. UPV-EHU. 1993.

Artola X., Evrard F. Dictionnaire intelligent d'aide à la compréhension, *Actas IV Congreso Int. EURALEX´90* (Benalmádena), 45-57. Barcelona: Biblograph, 1992.

Collins A.M., Loftus E.F. "A spreading activation theory of semantic processing", *Psych. Rev.*, vol. 82, no. 9, Sept. 1975

Kukich K. Techniques for Automatically Correcting Words in Text, in *ACM Computing sureys*, vol. 24, no. 4. December 1992.

Quillian, M.R. Semantic Memory in M. Minsky ed., p. 227-270, *Semantic Information Processing*. Cambridge (Mass.): MIT Press, 1968.

Rada, R., Mili, H., Bicknell, E. and Blettner, M. Development and Applicarion od a Metric on Semantic Nets, in *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 19, no. 1, 17-30. 1989.

Sussna, M. Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network, in *Proceedings of the Second International Conference on Information and Knowledge Management*, Airlington, Virginia USA. 1993.

Velardi P., Fasolo M. and Pazienza M.T. How to encode Semantic Knowledge: a Method for Meaning Representation and Computer-Aided Acquisition, *Computational Linguistics* 17, 2. 1991.