Automatic Contexonym Organizing Model (ACOM)

Hyungsuk Ji (jihs@isc.cnrs.fr)
Institut National Polytechnique de Grenoble
Institut des Sciences Cognitives

Sabine Ploux (ploux@isc.cnrs.fr)
UMR 5015 CNRS – University of Lyon I
Institut des Sciences Cognitives
67, boulevard Pinel
69675 BRON cedex, France

Abstract

Normal language user's word-association intuition (e.g. *drunken – stagger*) raises questions about the mental lexicon organization and its application for natural language processing tasks. We present an automatic contextually related words (*contexonym*) organizing model (ACOM) that reflects this intuition, giving one of the possible answers to this question. Trained on large corpora, the model (1) selects contexonyms for a given word and (2) classifies these groups of related words on a geometric representation. Some near-synonyms discussed in *Near-Synonymy and Lexical Choice* (Edmonds and Hirst, 2002) were chosen to test the model. The results showed that our model provides valuable contexonyms that reflect different usage and nuance of each word. Furthermore, the test on polysemous words showed that the model can classify contexonyms by grouping the different senses of a target word. The model can can be used as both theoretical lexicon-related research and practical natural language processing (NLP) research as well as an interactive reference.

Introduction

For any given word, we can generate many others related to it. Hearing the word *snow* for example, we think of words like *winter, ski, cold, white*, etc. Similarly, if we encounter words *learn, teacher* and *school* followed by the word *pupil*, we would interpret *pupil* as 'student' rather than 'opening of the eye'. Of course this is not always true: for someone, *snow* might evoke a completely different set of words from the above depending on her/his individual experience; the previous interpretation of *pupil* are no longer valid for a sentence like "In school, the teacher was examining the pupil of a fainted student.".

This discrepant feature makes us think over the nature of word-association intuition. Clearly it depends much on individual linguistic and extra-linguistic experience. But is it so arbitrary, compared to synonymy or antonymy, to generalize?

Consider if an English speaker is asked to (1) give synonyms for an English verb that describes an unstable walk or (2) give verbs that describe a drunken man's unstable walk. Does the second task require much longer time than the first? Are the words answered in the second task less homogeneous compared to those in the first one? In fact most English speakers would choose *stagger* and possibly *reel* for the second task without hesitation. On the contrary, non-English speakers or machines would have considerable difficulty in performing such a task even with all available references [1]. This absence of appropriate reference for contextually related words – far from justifying the useless of their generalization – may imply that our understanding on the mental lexicon remains still immature.

Indeed, Edmonds and Hirst expressed the need for such references that can be used in their computational linguistic model in discussing fine-grained word senses(Edmonds and Hirst, 2002): stupidity, blameworthiness, pejorative attitude and higher concreteness for *blunder* vs. *error*; writing-related mistake *slip*; memory-related mistake *lapse*; larger and animal-or-hunt-related *forest* vs. smaller *wood* (Gove, 1984, Room, 1985 as cited in (Edmonds and Hirst, 2002)). While they rightly pointed out the importance of the fine-grained differences of near-synonyms, the problem on how to develop an automatic process without the aid of lexicographer-like experts was not addressed.

Concerning this rather practical problem, automatic generation of a decision list for the target word – though it was focused on word sense disambiguation – was proposed (Yarowsky, 1995). After iterative processes, decision lists like *car, union, equipment, assembly, nuclear, job, etc.* were obtained for *(industrial) plant*. Though different in nature, latent semantic analysis (LSA) also generated series of words related to the target word.

Yet, in both approaches, automatic classification is missing: identifying seed words step in Yarowsky's model needs human intervention and LSA, applied several spheres such as automatic text evaluation (Landauer et al., 1998), metaphor problem (Kintsch, 2003), lacks this automatic classification (Laham's categorization (Laham, 1997), using encyclopedia as a source text, is closer to matching task than general automatic classification).

[1] For example, WordNet suggests *careen, keel, lurch, reel, stagger* and *swag* as synonyms for an unstable walk but no specific indication for the usage in such a situation; the query *drunken+unstable+walk* in search engines would fail also.

A fully automatic sense discriminating method based on a second-order has been proposed (Schütze, 1998). This approach shares with LSA their indirect nature: unlike a direct method such as Yarowsky's, they take into account the relations of the whole words rather than to focus on the target word and its neighbors.

Although such an indirect approach proved to be effective for certain tasks such as document classification, it loses in fact precious advantages of a direct approach. For example, it does not produce explicit related words to a target words which are important output to be used in theoretical or practical research. These cue words make it possible to also have a more sophisticated human mimic agent (understanding atypical sentence, puns, etc).

As for a direct approach, since it focuses on the target word, it fails to consider complex effects that other words could make. That is, although the neighbors of the target word are globally checked, the neighbors of the words other than the target word are not seriously taken into account. In consequence, this approach fails to classify properly the obtained words set.

We present here a model that can automatically discriminate words' senses like indirect approaches, but without losing rich features of a direct approach. Furthermore, the model proposes dynamic and continual representation of a target word which reflects human language users' intuition.

The model uses the minimal senses of words (*cliques*) that was first proposed by Ploux et al. to represent words' different semantic values (Ploux, 1997; Ploux and Victorri, 1998). In their study, cliques were obtained from non-sense-classified synonym database, and they were used to organize and represent words' senses. An evolved model based on a mapping method was used to represent a two-languages synonym representation (Ploux and Ji, 2003). This is in a sense a response to the problem of arbitrary organization in conventional dictionaries as pointed out by Dolan (inability to represent semantic distance between defined senses and hence the failure of organizing the senses properly(Dolan, 1994)) among others (Fellbaum, 1998; Budanitsky and Hirst, 2001; Pustejovsky and Boguraev, 1994).

The main difference between the present model and the previous one is that the present model does not need any kind of hand-coded references. Moreover, different sets of related words and cliques can be obtained according to chosen criteria. This will be explained later.

Contexonym

We define contexonym as relevant contextually related words for a target word. Contexonym is not symmetric or transitive contrary to synonym or antonym (that is, when a target word W has contexonyms $c_1, c_2, ..., c_k$, W is not necessarily a contexonym of $c_i (1 \leq i \leq k)$ and is true between c_is). Second, unlike synonyms or antonyms, contexonyms are often mixed grammar categories. We hypothesize that if the more adequate a training corpus is, the more relevant and more robust the contexonyms obtained from it will be. By an adequate corpus, we mean a sufficiently large and (1) well balanced corpus or (2) specific one depending on the research focus.

The procedure for constructing an automatic contexonym-organizing model is briefly presented below.

Model

STEP 1

For a given corpus, co-occurrences of all words in a defined passage (a sentence in this study) are counted and stored. Each headword $W_i^n (1 \leq i \leq N$, where N is the total number of the headwords in the database) has the whole types that co-occurred with it in a sentence; and each child c_j is arranged in descending order of co-occurrence with W_i^n:

$$W_i^n : c_1, c_2, \ldots, c_n.$$

STEP 2

For the target word, word-association table is constructed using four factors.

STEP 2-1 The first α portion (where $\alpha (0 < \alpha \leq 1)$) of the words (i.e. children that frequently co-occur with W_i^n) is selected and W_i^n becomes:

$$W_i^n : c_1, c_2, \ldots, c_{k\cdot},$$

where $k = n\alpha$ and n is the original number of children of W_i^n.

STEP 2-2 Similarly, the factor $\beta (0 < \beta \leq 1)$ serves to cut off rarely co-occurring children of the child c_j:

$$c_j^m : g_1, g_2, \ldots, g_l \quad (1 \leq j \leq k, \quad l = m\beta).$$

In this way, the following word-association table is obtained.

Table 1: Candadate contexonym table.

Headword	Selected	Rejected
W_i^n	$c_1, c_2, \ldots, c_k$	$c_{k+1}, \ldots, c_n$
c_1^m	$g_1, g_2, \ldots, g_l$	$g_{l+1}, \ldots, g_m$
...		
c_k^p	$h_1, h_2, \ldots, h_q$	$h_{q+1}, \ldots, h_p$

STEP 2-3 The factor $\gamma (0 < \gamma \leq 1, \gamma \leq \beta)$ has the same role as β except that γ is smaller. This gives another word-association table (Table 2) which will be used later to obtain cliques.

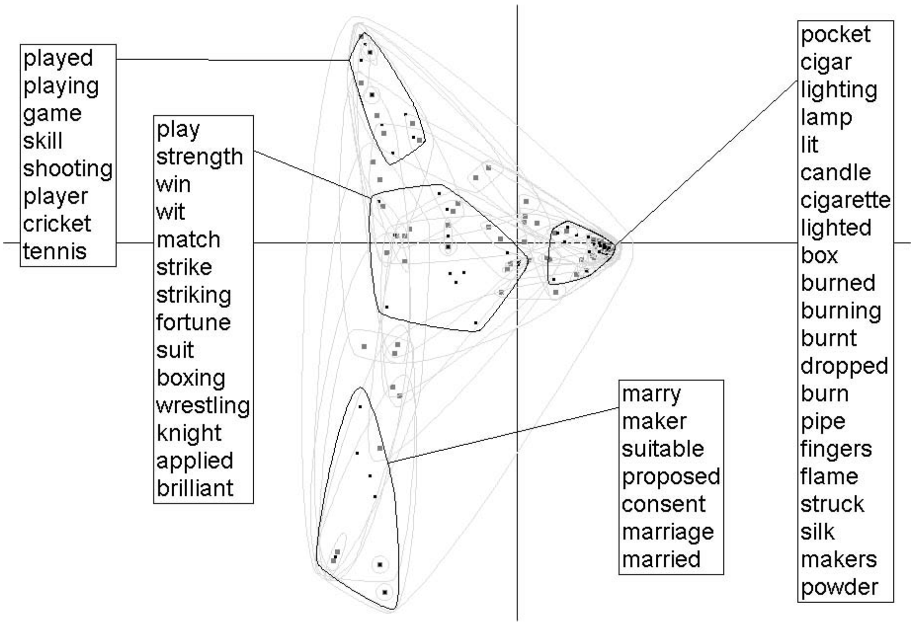

Figure 1: Representation of *match* in a semantic space ($\alpha = \beta = \gamma = 0.10$).
The gray pixel indicates *clique* and the black one shows the central mass of the cliques that belong to a contexonym. The clustering was carried out using the central mass of the cliques. This output is the projection on the principal plane of a multi-dimensional space.

Table 2: Second contexonym table.

Headword	Selected	Rejected
W_i^n	$c_1, c_2, \ldots, c_k$	$c_{k+1}, \ldots, c_n$
c_1^m	$g_1, g_2, \ldots, g_{l'}$	$\ldots, g_l, \ldots, g_m$
$\ldots$		
c_k^p	$h_1, h_2, \ldots, h_{q'}$	$\ldots, h_q, \ldots, h_p$

STEP 2-4 The factor δ is on/off Boolean. If the headword W_i^n is not found among c_j children $(g_1, \ldots, g_l)$ in Table 1, c_j itself in W_i^n and the c_j row (which contains c_j's children) are removed from both tables whenever δ is on. This filtering step gives the following final contexonym set (C_i^n) for W_i^n:

$$C_i^n = \{c_i : 1 \leq i \leq k, c_i \notin D\} \quad (k = n\alpha),$$

where D is the set of c_j words removed by filtering.

STEP 3

Cliques are calculated from these two tables. A clique is a mathematical term in graph theory meaning a maximum, complete subgraph. If w_1 has w_2 and w_3 as its member and vice versa for w_2 and w_3, then w_1, w_2 and w_3 form a clique. Otherwise, if say w_3 has only w_1 as its member, they fail to form a clique. If w_1, w_2, w_3, and w_4 form another clique, it *absorb* the clique w_1, w_2, w_3, resulting in only one clique. Table 2 can be used to calculate these cliques. Composed of several sets of words, cliques are considered in our model as 'minimal unit of a contexonym' that represent finer meanings than the word itself.

STEP 4

A correspondence factor analysis (proposed by Benzécri (Benzécri, 1992)) was used to represent correlations between cliques. The output is represented as a geometric semantic space that has as many axes as the total number of contexonyms chosen, in such a way that each axis could represent the corresponding word. Since every clique has its own coordinate, clique distances are proportional to their relatedness. The distance χ^2 between two cliques, y_i and y_j, is calculated in order to represent the cliques in a multi-dimensional space:

$$\chi^2(y_i, y_j) = \sum_{k=1}^{n} \frac{x_{..}}{x_{.k}} \left(\frac{x_{ik}}{x_{i.}} - \frac{x_{jk}}{x_{j.}} \right)^2,$$

where $x_{..} = \sum_{i=1}^{n} \sum_{j=1}^{p} x_{ji}$ and $x_{.i} = \sum_{k=1}^{p} x_{ki}$, $x_{i.} = \sum_{k=1}^{n} x_{ik}$; n is the total number of contexonyms and p is that of cliques; x_{ji} is equal to 1 if the i^{th} contexonym belongs to the j^{th} clique, and equal to 0 otherwise.

When (1) a clique (y_i or y_j) has many contexonym members or (2) many contexonyms belong to cliques

Word	α β γ	Contexonyms
blunder	0.07 0.15 0.05	[commit, committed, mistake] [stupid]
	0.35 0.50 0.35	{blunder, mistake, corrected, unpardonable, commit, committed, grievous, fatal, frightful, mistakes, repair} {stupidity, gross, stupid} {joke, unlucky}
	0.50 0.50 0.50	{stupidity, commits, gross, stupid, blunders, corrected, detected, mistake, blunder, repair} {indiscretion, unpardonable, committing, commit, mistakes, grievous, committed, fatal} {calamity, frightful} {awkward, joke, unlucky} [howells]
error	0.07 0.15 0.05	{argument, opinions, belief, faith, contrary, knowledge, source, greater, taught, easily, imagine, liberty, due, divine, former, understanding, experience, regard, merely, appears, authority} {error, admit, opinion, convinced, correct} {discovered, ideas, political, principle, causes, doctrine, degree, mere, religion, science, fundamental, modern, discover, method} {false, judgment, evil, virtue, conduct, ignorance, judge, lead, wise, fallen, makes, avoid, fall} {prove, wrong, errors, liable} {vulgar, sin, ways, commit, trial, fault, guilty, committed, lies} {grave, mistaken, , corrected, mistake, supposing} {serious, fatal} [clerical] [pointed] [text]
lapse	0.07 0.15 0.05	[{lapse, changes, slow} {centuries, ages, century} {geological, strata}] [memory]
	0.10 0.25 0.10	{strata, geological, organic, centuries, ages, century, lapse, slow, changes, species, rate, period, progress} {forgotten, memory} {minutes, interval, weeks, absence, months} {moments, recall}
slip	0.07 0.15 0.05	{foot, narrow, slip, fingers, hole, corner, quietly, coat, watch} {advantage, allowed, chance, letting, opportunity, reach, easily, easy, managed, try, lest, fall, caused, escape} {written, handed, wrote, tongue, pocket, paper, inside, lines} {fast, noose, rope, boat, knot, neck}
mistake	0.05 0.10 0.05	{discovered, wrong, error, instead, fatal, committed, supposing, mistake, serious} {impossible, correct, makes} {thinking, meaning, perceived} {possibility}
	0.07 0.15 0.05	[{instead, mistake, opinion, mere, possibility, seeing, imagine, likely, unless, particular, supposed, discovered, easily, makes, easy, convinced, impossible} {slight, due, greatest, regard, greater, result, correct, false, giving, caused, committed, serious} {supposing, surely, error, marriage, possibly, mistaken, wrong, anybody, perceived, thinking} {intended, trying, meaning, explain, meant} {fatal, sad, fallen, terrible}] [sorry]
enjoined	0.10 0.25 0.10	{secrecy, silence, multitude, commanded, enjoined} {priests, penance, perform} {instructions, obedience} [strictly]
	0.50 0.50 0.50	{strictly, instructions, obedience, code, priestly, piety, enjoined, silence} {prohibited, positively, expressly, forbidden, obey, priests, penance, punishment, penalty, commanded, commands, executed, execute, perform, governors, whatsoever} {prudence, advised, abstain, multitude} {injunction, observance, strict, despatch, injunctions, caution, secrecy, strictest} {commended, earnestly, solemnly} [directors]
prescribed	0.07 0.15 0.05	{prescribed, regulations, rules, terms} {assembly, authority, duties, oath, thereof, provided, conditions, according, required, constitution, laws} {section} {mode, treatment} {patient, physician, medicine, medicines} {remedies, remedy} [limits]
ordered	0.05 0.05 0.05	{officer, officers, orders, royal, report, prisoner, prisoners, duke, ships, charge, emperor, appointed, governor, majesty, carry, ship, palace, send, arrived, horses, accordingly, follow, placed, immediately, ordered} {lieutenant, command, guard, commanded, join, camp, army, troops, move, division, regiment, soldiers, battle, enemy, advance, attack, brigade, cavalry, march, colonel, corps} {informed, board, finding, council, remain, receive, post, refused} {thrown, carriage, twelve, proceed, instantly, drive, paid} {clothes, servant, clock, servants} {bell, prepared, bottle, breakfast, dinner} {coffee, wine, supper, tea}
forest	0.03 0.04 0.03	{(wind, woods, green, fields, forest, covered, trees, depths, snow, blue, sky) (narrow, village, hill, stream, broad, below, wide) (mountains, hills, mountain, plain, valley, distance, mile) (field, grass, wood, beasts, dense, beneath, tree) (east, vast, shore, west, lake, north, distant, plains) (oak, grew, leaves, birds, flowers, forth, tall, branches, thick, pine, summer, golden, spring) (path, edge, foot)} {deer, hunting} {knight, castle, rode}
woods	0.03 0.04 0.03	{snow, pine, sky, mountain, hills, meadows, forest, trees, woods} {leaves, flowers, bird, birds, green, tree, grass, fields, wood, winter, blue, gray, yellow, spring, thick, golden, summer} {rivers, streams, lake, valley, distant, rocks, forests, places, mountains, waters} {mile, shore, hill, edge, stream} {path, walk, walking}
drunken	0.07 0.15 0.05	{wine, drunk, sober, drink, drinking, sleep, singing, drunken, laughter, song} {fool, mad} {dirty, streets} [brute] [reeled, staggered] [reeling, staggering] [sailor]
	0.10 0.25 0.10	[{drunken, mad, bar, fallen, drink, fit, asleep, sleep, devil, lot, fool, worse, dog, legs, dirty, trying} {sailor, dancing, singing, sailors, songs, crying, laughter, dance, soldiers, shouts, cries, streets} {eaten, liquor, sober, feast, drank, swearing, drunk, drinking, wine, soldier, song} {brutal, creature, brute, coarse} {crowd, fury, mob, riot}] [reeled, reeling, staggered, staggering] [stagger]

Table 3: Output of test on Edmonds and Hirst's examples and *drunken*.

y_i and y_j, they should be less representative. This was considered in the first and second terms of the equation, respectively, by a distance-reducing effect.

STEP 5

Cliques are projected onto a two dimensional space. The center of mass of the cliques which belong to a contexonym corresponds to this contexonym. Either cliques or this contexonyms is grouped to form a few clusters using hierarchical clustering method. As will be discussed later, this cluster should be considered as rough boundary rather than absolute class. The advantage of the geometric representation is that it represents continuous minute change of the relatedness between contexonyms.

Figure 1 shows the result obtained for the word *match* after the model was trained on an English corpus with 240 million words. In this figure, four major senses are successfully distinguished. Note that, any kind of the electronic dictionary or encyclopedia was used to train the model.

Experiments

Test on Edmonds and Hirst's Examples

The model was trained on an English corpus maintained by Project Gutenberg, which includes literatures, essays, and other writings. Any kind of electronic dictionaries or encyclopedia was excluded from the train corpus because they are already hand-coded references and our main goal is to construct automatically relevant sets of contexonyms from non-knowledge structured texts. The total number of tokens in the training corpus counts more than 240 million.

The near-synonyms which were carefully investigated by Edmonds and Hirst were tested.

As shown in Table 3, while *blunder* has the contexonyms *stupid* and stupidity, there are no such contexonyms for *error*, suggesting that the former has 'stupidity' as a connotation. Contexonyms like *calamity, frightful, fatal, grievous*; *awkwark*; *indiscretion, unpardonable* characterize the target word *blunder* by its 'strength', 'blameworthiness', 'pejorativity', which distinguish the word from *error.* On the other hand, contexonyms like *discovered, ideas, political, principle, causes, doctrine, religion, science, fundamental, modern, discover* and *method* of *error* suggest that this word is used in scientific and political contexts.

The contexonyms of *lapse* like *forgotten, memory, minutes*, and *weeks* also reflect the word's usage. Among other senses of the word *slip*, *written, handed, wrote, lines*, and *tongue* suggest its usage in speech(or writing)-related mistakes.

The contexonyms of *woods* like *houses, path, walk*, and *walking* contrast with those of *forest* like *deer, beasts, hunting, castle*, and *knight.* This is consistent with Room's observation(1985, as cited in (Edmonds and Hirst, 2002)).

Discussion

In this paper, we presented a model that automatically produces and organizes contexonyms for a target word. The test results show that the model is able to classify contexonyms as well as to reflect words' minute usage and nuance. In addition, what the model reflects can be extended to broader knowledge representation such as historical one (e.g. *Egypt* [2]) or situational one (e.g. (*actor, concert, curtain, opera, performance, play, spectators, ticket* for *theater*).

The model also shows automatic evolving features. For example, after trained on the French newspaper *Le Monde*, the model generated, for *vache (cow)*, the contexonyms *folle (mad), ESB(BSE), embargo, Spongiforme, Creutzfeldt, Jakob, etc*, reflecting recent mad cow issue in Europe [3].

This automatic feature of the model has some advantages over a manual coding approach. First, some usages of a word apt to miss to compile are easily captured by our method. Scientific usage of *error* discussed above is one of such examples.

Second, rapidly changing issues, which are too wide to be coded manually, can be updated by automatic approach. The most widely used machine translator like Systran, Babel Fish, and FreeTranslation interpreted the word *match* as a *wooden lighter* and wrongly translated the word into the French word *allumette* in the sentence that includes significant cue words such as *final, Sampras, wins, champions*, and *Agassi* [4]. Trained on *Le Monde*, the model generate relevant cue words [5], suggesting sports-

[2] *egypt* ($\alpha = \beta = \gamma = 0.05$) { cities, nations, rome, greek, roman, kings, ancient, kingdom, africa, cyprus, palestine, cambyses, syria } { abraham, israel, sons, egyptians, priests, temple, sacred, thence, jews, worship, babylon, jerusalem, alexander, gods, caesar, greeks, egyptian, temples, bonaparte, egypt, ptolemy, alexandria, cleopatra } { china, india, arabia, persia, countries, asia, europe, civilization, italy, conquered, empire, greece } { brethren, jacob, exodus, joseph, pharaoh, israelites, moses } { slave, cairo, upper, expedition, throne, pyramids, desert, nile } { al, sultan }

[3] *vache* ($\alpha = \beta = \gamma = 0.02$) { bovine, alimentation, mesures, farines, agriculture, animaux, animales, animale, bovins, contamination, britanniques, sang, esb, maladie, alimentaire, agricole, bretagne, folle, vache } { embargo, interdiction, boeuf, viande } { transmission, agent, humaine, pathie, spongiforme, creutzfeldt, jakob } { afssa, aliments, experts, sanitaire } { lait }

[4] The final was Hewitt's first and Sampras' 17th, but the less experienced 20-year-old Australian was much more energetic. After consecutive wins against former champions Pat Rafter, Andre Agassi and Marat Safin, Sampras appeared to have nothing left for his second match in barely 24 hours.

[5] *Agassi+Sampras* ($\alpha = \beta = \gamma = 0.05$) { terre, mondial, tennis, agassi, andre, roland, patrick, australie, demi } { rafter, chelem, wimbledon, australien, open,

related contexts. The pre-calculation-free feature (unlike LSA or Schütze's model) of the model makes it easier to evolve by simply adding newly created database to an existing one.

The model presented here can be used as a reference for lexicographers or foreign language learners. On-line users could test all types of words (more than 200,000) in the corpus (http://dico.isc.cnrs.fr/dico/context/search) and may obtain visual representation like Figure 1. Besides this practical usage, the model could be used for a theoretical research on the mental lexicon by inspiring possible mechanism or by simulating theoretical results.

References

Benzécri, J.-P. (1992). *Correspondence Analysis Handbook*. Marcel Dekker, New York.

Budanitsky, A. and Hirst, G. (2001). Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the ACL*, Pittsburgh, PA.

Dolan, W. B. (1994). Word sense ambiguation: clustering related senses. In *Proceedings of COLING94*, pages 712–716.

Edmonds, P. and Hirst, G. (2002). Near-synonymy and lexical choice. *Computational Linguistics*, 28(2):105–144.

Fellbaum, C. D. (1998). *WordNet: An Electronic Lexical Database*. MIT Press, New York.

Kintsch, W. (2003). Metaphor comprehension: A computational theory. *Psychonomic Bulletin & Review*. In press.

Laham, D. (1997). Latent semantic analysis approaches to categorization. In Shafto, M. G. and Langley, P., editors, *Proceedings of the 19th annual meeting of the Cognitive Science Society*, page 979, Mawhwah, NJ. Erlbaum.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25:259–284.

Ploux, S. (1997). Modélisation et traitement informatique de la synonymie. *Linguisticæ Investigationes*, XXI(1):1–28.

Ploux, S. and Ji, H. (2003). A model for matching semantic maps between languages (French / English, English / French). *Computational Linguistics*, 29(2):155–178.

Ploux, S. and Victorri, B. (1998). Construction d'espaces sémantiques à l'aide de dictionnaires informatisés des synonymes. *TAL*, 39(1):161–182.

Pustejovsky, J. and Boguraev, B. (1994). Lexical knowledge representation and natural language processing. *Artificial Intelligence*, 63:193–223.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

battu, joueur, vainqueur, finale, tournoi, no, sampras, garros, internationaux, pete } { martin, arnaud, michael } { gustavo, kuerten } { tenant }

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the ACL*, pages 189–196. Cambridge, MA.