# Lexical Knowledge Representation with Contexonyms

**Hyungsuk Ji**
ISC*
INPG†
jihs@isc.cnrs.fr

**Sabine Ploux**
ISC
University of Lyon I
ploux@isc.cnrs.fr

**Eric Wehrli**
LATL‡
University of Geneva
wehrli@latl.unige.ch

## Abstract

Inter-word associations like *stagger - drunken*, or intra-word sense divisions (e.g. *write a diary* vs. *write an article*) are difficult to compile using a traditional lexicographic approach. As an alternative, we present a model that reflects this kind of subtle lexical knowledge. Based on the minimal sense of a word (*clique*), the model (1) selects contextually related words (*contexonyms*) and (2) classifies them in a multi-dimensional semantic space. Trained on very large corpora, the model provides relevant, organized contexonyms that reflect the fine-grained connotations and contextual usage of the target word, as well as the distinct senses of homonyms and polysemous words. Further study on the neighbor effect showed that the model can handle the data sparseness problem.

## 1 Introduction

With progress in natural language processing techniques (NLP), increasingly sophisticated models and methods have been proposed in the machine translation (MT) research. New techniques distinguish the minute differences between similar words (Edmonds and Hirst, 2002) or take into account collocations (Edmonds, 1997), idioms (Wehrli, 1998), or contextually related words (Dagan and Itai, 1994; Lin and Pantel, 2002), etc.

This kind of approach depends to varying extents on adequate references. For the fine-grained lexical knowledge model (FLK) (Edmonds and Hirst, 2002), having adequate references is indispensable, or the model will not work in practical applications.

However, such detailed references are limited in number, and manual lexicographic coding is too time-consuming to continuously update new information. Other problems with the classical lexicographic organization have been pointed out, such as its inability to represent the semantic distance between defined senses and its failure to properly organize the senses, and alternatives have been proposed (Dolan, 1994; Budanitsky and Hirst, 2001; Fellbaum, 1998; Manning, 1993; Ploux, 1997; Pustejovsky and Boguraev, 1994).

In addition, subtle lexical knowledge is too vague and too broad to handle. For instance, relations like *stagger - drunken*, which could be informative for non-English speakers or machines, are too numerous to be processed. Intra-word relations share this problem: while the English word *write* is considered to have the same semantic value in "*write a diary*" and "*write an article*", the French words *écrire* and *rédiger*, respectively, are widely used in these two phrases. This sort of sense division is also too minute and too frequent to be captured using conventional manual lexicography techniques.

An alternative would therefore be to automatically generate the related words for a given word, which could serve as a reference. Clearly, contextually related words are meaningful indicators of the target word's semantic value in a given context. For instance, two sets of words { lit, candle, cigarette } and { tennis, final, win } are trustworthy cue-word sets for disambiguating the word *match*; *stupid* is more closely related to *blunder* than to *error* (Edmonds and Hirst, 2002), and *peace* distinguishes *treaty* from *contract* (Dagan and Itai, 1994).

Such word lists may be obtained for target words by selecting seed words and performing an iterative, decision-list-making task (Yarowsky, 1995), or by latent semantic indexing (LSI) (Landauer et al., 1998). A common limitation of these approaches, however, is that they do not provide a fully automatic method for organizing the related words obtained: identifying seed words needs human intervention and LSI does not provide an automatic classification other than a restricted matching-based one that requires an encyclopedia as a source text (Laham, 1997).

*Institut des Sciences Cognitives, UMR 5015 CNRS, 67, boulevard Pinel, 69675 BRON cedex France
†Institut National Polytechnique de Grenoble
‡Laboratoire d'Analyse et de Technologie du Langage

A fully automated sense-discrimination method based on a second-order comparison in semantic space has been proposed (Schütze, 1998). Because this approach focuses on comparing vectors for disambiguation, it does not explicitly produce a relevant set of words. Unlike a direct method (e.g. Yarowsky's), this technique takes all word relations into account, not just those between the target word and its neighbors. This technique proved effective for data sparseness problems (along with LSI) but it has some distance to go for lexical knowledge representation. For instance, words that have never co-occurred with a target word can, in principle, be the closest ones to it.

Dagan and Itai demonstrated how contextually related words could contribute to selecting a proper target word in MT tasks (Dagan and Itai, 1994). However, since their study focused on target-word selection, the problem of how to organize and assign source-word senses was not addressed.

In this paper, we present a model that explicitly produces contextually related words and classifies them after training on a large corpus. The model uses a rather straightforward method, in the sense that it considers co-occurrences of words. The main distinction between the model presented here and other statistical ones is that it generates the minimal senses of words (*cliques*) in order to organize the related words obtained. Cliques are then represented on the principal plane. This makes it possible to represent several target words in MT tasks.

Ploux et al. proposed the prototype of a model that represents synonym senses from a non-sense-classified synonym database (Ploux, 1997; Ploux and Victorri, 1998) and a two-language synonym-matching model based on a mapping method (Ploux and Ji, 2003). The main difference between the present model and the previous one is that the present model is fully automated, insofar as it does not need any kind of hand-coded references, only raw text sources. Furthermore, different sets of cliques can be obtained according to chosen criteria. This will be explained later.

## 2 Contexonym

We define contexonyms as relevant contextually related words for a target word. By context, we mean a certain number of neighboring words of the target word (from a small-sized window to one or more paragraphs). Unlike synonyms or antonyms, contexonyms are not symmetric or transitive (i.e., when target word $W$ has contexonyms $c_1, c_2, ..., c_n$, $W$ is not necessarily a contexonym of $c_i (1 \leq i \leq k)$, and this is also true between $c_i$s).

Second, contexonyms are more dynamic and fuzzy than synonyms or antonyms, and they evolve faster (cf. Ji and Ploux, 2003). Without any neighboring words, the contexonyms of a target word reflect various typical semantic relations. The word *match*, for example, may have contexonyms related to "wooden lighter", "game" and "marriage". On the other hand, the presence of the neighbors of the word could change this equivocal situation, giving a polarized semantic field of the word. Finally, unlike synonyms or antonyms, contexonyms are often from mixed grammar categories.

We hypothesize that the more adequate a training corpus is, the more relevant and robust the contexonyms obtained from it will be. By an adequate corpus, we mean a sufficiently large and well balanced corpus.[1]. If correctly designed, the model is expected to also simulate the above characteristics of contexonyms.

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

## 3 Model

### STEP 1

For a given corpus, co-occurrences of all types 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 types in the corpus) has children ($c_j$s) that are arranged in descending order of co-occurrence with $W_i^n$; children with co-occurrences smaller than a 10,000th of the global frequency of the headword $W_i^n$ are removed to reduce noise:

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

[1] A specific corpus would be considered adequate if a specific domain is being processed (e.g. science, religion, or spoken language).

**STEP 2**

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

**STEP 2-1**

In order to eliminate children that rarely co-occur with $W_i^n$, the first $\alpha$ portion (where $0 < \alpha \leq 1$)) of the words is selected. And $W_i^n$ becomes:

$$W_i^n : c_1, c_2, \ldots, c_k,$$

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

**STEP 2-2**

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:

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

Table 1: Candidate contexonym table.

**STEP 2-3**

The factor $\gamma(0 < \gamma \leq \beta)$ has the same role as $\beta$ except that $\gamma$ is smaller, which makes it possible to have different sets of cliques without changing the global contexonyms obtained in the previous step. This gives another word-association table (Table 2), which will be used later to obtain cliques ($l' = m\gamma$ and $q' = p\gamma$).

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

Table 2: Second contexonym table.

**STEP 2-4**

The factor $\delta$ 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 $\delta$ is on (in this study, $\delta$ was set to 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 members 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 'absorbs' 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 units 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. The distance $\chi^2$ 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 $x_{..} = \sum_{i=1}^{n} \sum_{j=1}^{p} x_{ji}$, $x_{.i} = \sum_{k=1}^{p} x_{ki}$ and $x_{i.} = \sum_{k=1}^{n} x_{ik}$; $n$ is the total number of contexonyms and $p$ is the total number of cliques; $x_{ji}$ is equal to 1 if the $i^{th}$ contexonym belongs to the $j^{th}$ clique, and equal to 0 otherwise. Since every clique

has its own coordinates, clique distances are proportional to clique relatedness.

When (1) cliques $y_i$ and $y_j$ have many contexonym members or (2) many contexonyms belong to cliques $y_i$ 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 and are classified by hierarchical clustering. This detailed feature of the model is explained with some examples below.

## 4 Test on Examples

The model was first trained on an English corpus maintained by Project Gutenberg (PG), which includes literature, essays, and other writings. Any kind of electronic dictionary or encyclopedia was excluded from the training corpus. The database thus constructed was combined with another, separate database trained on the British National Corpus (BNC). The total number of tokens in the training corpora was over 300 million. For French words, the model was trained on five years of the French newspapers *Le Monde* and *L'Humanité*.

With $\alpha = \beta = \gamma = 0.05$, 50 contexonyms and 133 cliques were obtained for the target word *match*. Some of the cliques are:

- 1: applied, marriage, match, proved
  6: box, candle, dropped, lighted, match, struck, threw
  68: burned, candle, flame, lamp, lighted, lit, match
  93: cigar, cigarette, lighted, lit, match, pipe
  109: fight, game, match, proved, shot, won

While the contexonyms *shooting* and *maker* each belong to only one clique, *struck* belongs to 49 cliques. In other words, *maker* has only one minimal semantic value and *struck* has 49 'different semantic values'. This difference is represented in Figure 1 by the region that each contexonym covers (i.e., possessing clique points).

Clustering can be done with either cliques or contexonyms. In this study, the latter was always used.

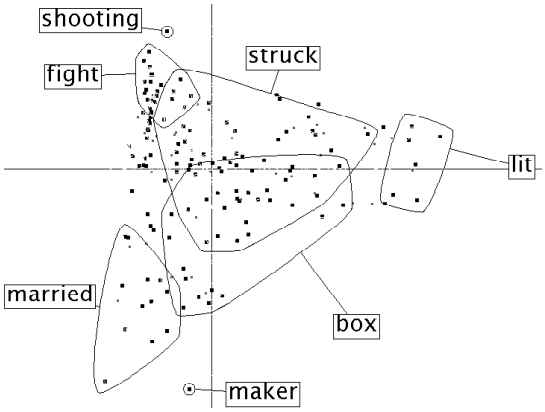


Figure 1: Some contexonyms of *match*.

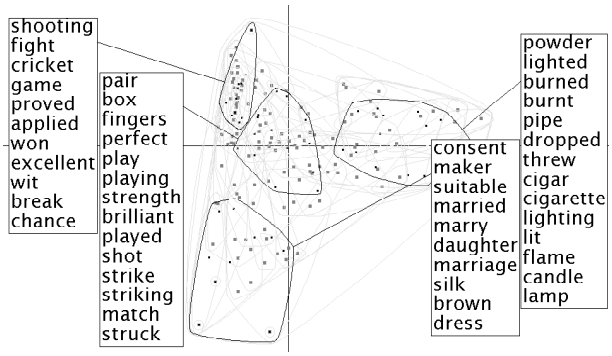


Figure 2: Classification of the contexonyms of *match*.

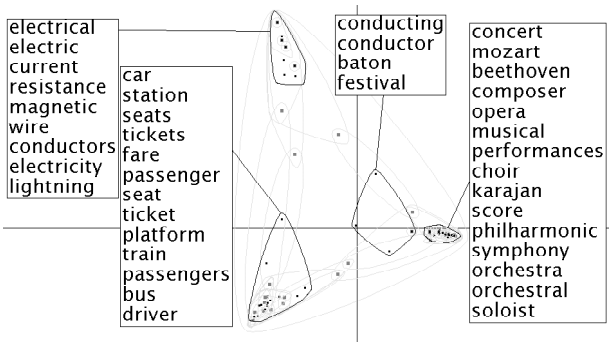


Figure 3: Representation of *conductor*.

| Word | $\alpha$ | $\beta$ | $\gamma$ | Contexonyms |
|---|---|---|---|---|
| blunder | 0.05 | 0.10 | 0.05 | {blunder, mistake} {commit, committed} {stupid} |
| | 0.30 | 0.30 | 0.30 | [{corrected, political, reckon, blunder, mistake, minister, serious, fatal, pardon, terrible, awful, joke} {unpardonable, gross, committing, stupidity, grievous, stupid, ignorance, guilty, commit, committed, excuse, mistakes} {survived, tragic} {egregious}] [speelman] [tactical] |
| lapse | 0.10 | 0.20 | 0.10 | {considerable, mere, ten, twenty, lapse, slow, absence, rate, sudden, fifty, forgotten, memory, allowed, minute, months} {evidence, species, progress, century, original, changes, vast, ages, centuries} {recall, moments, interval, minutes} {geological, organic} {strata} {momentary} |
| slip | 0.05 | 0.10 | 0.05 | {past, slip, tried, run, fall, hold, managed, opportunity, try, allowed, easy, chance, easily} {tree, watch, rope, letting, quietly} {foot, caught, front, drew, fingers, neck, handed, pocket} {tongue, book, paper, written, wrote} |
| enjoined | 0.10 | 0.20 | 0.10 | {commands, strictly, obedience, instructions, obey, strict, commanded, enjoined, abstain, duty, expressly, orders} {multitude, earnestly, silence} {penance, perform, priests} {secrecy} |
| ordered | 0.05 | 0.05 | 0.05 | {captain, horse, immediately, company, send, placed, ready, six, service, ordered, party, court, pay} {troops, attack, enemy, city, command, line, horses, officers, war, army, soldiers} {carriage, dinner, master, table} {march} {costs} |
| error | 0.05 | 0.05 | 0.05 | {trial, evil, human, errors, lead, false, wrong, error, truth, correct, ignorance, fatal, committed, opinion, due, judgment, causes, serious, avoid, fault} {fallen, source, lies, ways, discovered, common, fall} {mistake} {mistaken, supposing} |

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

Figure 2 shows the output of this classification. Figure 3 is another example of such a representation[2].

In general, stricter constraints (smaller values of $\alpha, \beta$ and $\gamma$) give fewer contexoyms than lenient ones. Below are some examples. Brackets indicate disjoint relations between contexonyms, curly brackets denote classifications on the principal plane, and parentheses, classifications on a non-principal plane. The contexonyms *kick* and *kicked* for the word *bucket* suggest the idiom "*kick the bucket*" (an example in Wehrli, 1998), and the contexonym *article* for the French *rédiger* reflects their relatedness.

- **drunken** ($\alpha = 0.05 \quad \beta = 0.05 \quad \gamma = 0.05$): [brawl] [brute] [drink, drunk, sober, wine] [reeled, staggered] [reeling, staggering] [sailor] [stupor]

- **drunken** ($\alpha = 0.05 \quad \beta = 0.10 \quad \gamma = 0.05$): [ {(coarse, shouts, street, streets, dirty, songs, mad, fellow, driver, driving, dancing, fool, singing, drunken, laughter) (drunk, brute, asleep, wine, drank, sober, drink, song) (stupor, sailor, sailors, brawl, crying) (killed) (fury)} {eaten, feast} {brutal, mob} {reeled, staggered} {reeling, staggering}] [orgies] [stagger] [swearing]

- **bucket** ($\alpha = 0.05 \quad \beta = 0.05 \quad \gamma = 0.05$): [buckets] [contents, emptied, tin] [mop] [rope]

- **bucket** ($\alpha = 0.05 \quad \beta = 0.10 \quad \gamma = 0.05$): [{carrying, empty, filled, tin, wash, buckets, fill, contents, emptied, poured} {bottom, mop, bucket, ice, wooden} {chain, rope} {file, record}] [dipped] [kick, kicked] [packing] [pump] [spade]

- **rédigé** ($\alpha = 0.05 \quad \beta = 0.10 \quad \gamma = 0.05$): {lui-même, rendu, rédigé, signé, acte, article, présenté, publication, daté, document, base, instruction, chambre, avis, bureau, communiqué} {intitulé, lire, professeur, commun, essentiel, pages, publié, guide, manifeste, ouvrage, langue, mémoire, rapports} {code} {tiberi} {alinéa}

Since there is no room in this paper to list complete examples, our demonstration of the model is limited to the examples introduced in two articles that share some interests with the current paper. More than 100,000 types of

[2]From now on, no more than 30 of the most closely-related contexonyms are presented for each example (in lowercase).

| | |
|---|---|
| **peace** | {army, france, peace, war, law, nation, city, sense, free, cause, desire} {earth, happiness, live, spirit, hold, land} {foreign, nations, terms, justice, states, united} {heaven, happy, soul, joy, quiet} {security, **treaty**} |
| **treaty** | {terms, rights, united, states, treaty, france, french, british, foreign, spain, **peace**, powers, agreed, subject, war, article, britain} {alliance, majesty, concluded, emperor} {territory, commerce, american, citizens, congress, mexico} {**signed**, ratified, senate} |
| **chances** | {calculated, finding, risk, probability, ten, election, favour, nine, success, chances, favor} {game, missed, desperate, plenty, calculate, chance, getting, happen, escape, survival} {**increase**, increased, reduce, diminished, improve} {promotion, victory, winning} |
| **achieved** | {achieved, real, period, process, history, result, increase, development, greater, actually, considerable, **progress**, growth, results, purpose, per, objectives, using} {economic, aim, position, areas, ways} {success, record, sales, task} {independence, status} {victory} |
| **diplomats** | {libya, embassy, p., **talks**, sanctions, saudi, jan., libyan, reported, iraqi, kuwait, diplomatic, embassies, diplomats, ambassador, foreign} {politicians, journalists, consuls, countries, officials, intelligence, claimed, ministry} {expulsion, expelled, yugoslav} {diplomacy, eighteenth} {statesmen} |

Table 4: Test on Dagan and Itai's Examples

other words can be tested interactively on-line (http://dico.isc.cnrs.fr/dico/context/search).

### 4.1 Test on Edmonds and Hirst's Examples

In discussing near-synonymy, Edmonds and Hirst carefully investigated the subtle differences between the words *blunder*, *error*, *lapse* and *slip*, and the pairs of words *order*/*enjoin*, *forest*/*woods* (Edmonds and Hirst, 2002).

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 while the latter does not. Contexonyms like *unpardonable, fatal, grievous, awful, indiscretion* and *egregious* characterize the target word *blunder* by its strength, blameworthiness, and pejorative character, unlike the word *error*. The contexonyms of *lapse* like *forgotten, memory*, and *minutes* also reflect the word's usage; the contexonyms *written, wrote, lines* and *tongue*, among other senses of the word *slip*, suggest that it is used for mistakes in speech or writing.

The test on *woods* gave the contexonyms *houses, path, walk,* and *walking*, which were not among the contexonyms of *forest*, while *deer, beasts, hunting, castle* and *knight* were the contexonyms of *forest* and not of *woods*. This is consistent with Room's observation (1985, as cited in (Edmonds and Hirst, 2002)).

Overall, the fine-grained subcategorical differences between similar words, as discussed in the FLK model (Edmonds and Hirst, 2002), were successfully reflected here. Moreover, other subcategorical features that were not discussed in the original studies were found: the contexonyms of *error* reflect its scientific usage; the contexonyms *coffee, wine, supper* and *tea* for *order* suggest that the verb is applicable to asking for drinks ($\alpha = \beta = \gamma = 0.05$ for the GP corpus). This information is not trivial, since the English *order* in this situation should be translated into the French *commander* and not *ordonner*, which fits other situations such as military ones.

### 4.2 Test on Dagan and Itai's Examples

In discussing the problem of selecting a proper target word in MT, Dagan and Itai presented some examples: *sign* (rather than *seal*, *finish* or *close*) is the correct verb to use with *treaty*, and *treaty* (rather than *contract*) is the proper word to use with *peace*. In the following sentence, the first word in curly brackets is the correct one in each case (Dagan and Itai, 1994):

- Diplomats believe that the joining of Hon Son {increases | enlarges | magnifies} the chances for achieving {progress | advance | advance ment} in the {talks | conversations | calls}.

The five words *peace, treaty, chances, achieved* and *diplomats* were tested using the model trained on the GP and BNC corpora. In Table 4, the boldface words in the contexonym list are more closely related to the headword than similar words (e.g. *sealed, closed, enlarge, conversations,* etc.), which were not selected as contexonyms in the given factor condition.

| | |
|---|---|
| **match + wins** | matches, strength, play, chance, won, league, cricket, games, fight, club, final, game, points, united, record, victory, team, season, marriage, daughter, prix, match, box, draws, wins, loses, lighted, cigarette, lit, whoever |
| **match + wins + champions** | won, game, division, champions, defeat, final, united, games, champion, club, yesterday, season, draw, points, victory, strength, fight, play, team, players, australia, chance, defending, match, record, matches, wins, daughter, struck, lighted |
| **match + agassi + sampras** | tennis, champion, minutes, struck, play, final, davis, won, doubles, yesterday, players, michael, game, player, jim, tournament, today, defeat, victory, weeks, opening, monday, sets, agassi, match, wimbledon, edberg, goran, ivanisevic, sampras |
| **match + champions + wins + agassi + sampras** | final, champion, won, defeat, play, game, champions, doubles, player, league, players, tennis, davis, sets, minutes, michael, tournament, team, victory, season, yesterday, today, wins, weeks, jim, agassi, match, edberg, ivanisevic, sampras |

Table 5: The merging effect for match and its neighbors ($\alpha = \beta = \gamma = 0.05$). The contexonyms are ordered by nearness to the origin (i.e., representativeness).

### 4.3 Test with the Merging Method

Since the information carried by a target word's contexonyms is relevant, it is directly applicable to some MT tasks. Consider the sentence below:

- 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.

Widely-used machine translators like Systran, Babel Fish, and FreeTranslation incorrectly translate the words *the final* into the French words *le final*, and *match* into *allumette* (wooden lighter), whereas the correct translations are *la finale* and *match*, respectively.

Trained on five years of the French newspapers *Le Monde* and *L'Humanité*, our model produces the contexonyms *finale* and *match* for the target words *Agassi*, *champions* and *victoires*, and *finale* only for *Rafter* and *Sampras*. This clearly points to the correct target-language words.

Yet, two problems remain unsolved: first, unlike the target-language selection phase, no disambiguation is performed for the source language; second, potential data sparseness problems (Lee and Pereira, 1999) are not covered by this direct approach. The first problem involves assigning the meaning of the word *match* in the concerned paragraph to a proper cluster. But as Figure 2 shows, there are no contexonyms shared with the text in question.

We present the contexonym merging method as a remedy to these problems. This consists in inputting more than one target word to get the contexonyms. To compensate for the global frequency effect, a normalizing merging method is used. Table 5 shows the gradual exclusion of less-closely-related contexonyms like *cigarette* and *marriage*, and the gradual inclusion of more relevant ones such as *player* and *tennis*.

Another way to discriminate the target word's sense is to observe the *linking contexonyms*. For example, *match* and *wins* have no shared cliques and the region they cover is disjoint in the principal projection. But they are linked by the intermediate contexonyms *play* and *games*, which share areas with the *match* and *wins* cliques.

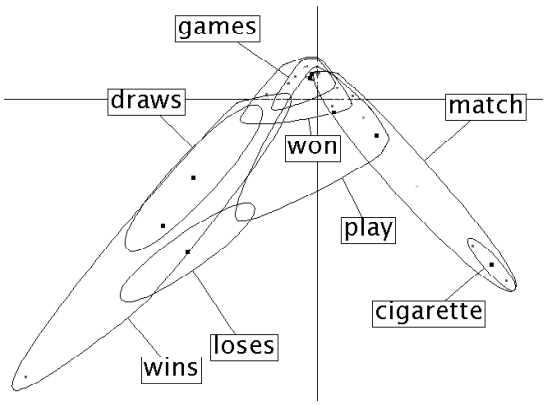


Figure 4: Merging of *wins* and *match*.

One solution to the data sparseness problem is to build a decision list, as proposed by Yarowsky (Yarowsky, 1995), using contexonyms as

starting seed words[3]. Although the current model was not designed to solve such a problem, the merging method could be considered as an indirect alternative. For instance, *stars* and *mathematician* have no common contexonyms but they are linked by *astronomer* in a merging search, suggesting that *stars* should be interpreted as celestial bodies and not as actors.

## 5 Conclusion

In this paper, we presented a model that automatically produces and organizes the contexonyms of a target word. The test results show that the model (1) reflects the fine-grained senses of the word, and (2) provides typical trustworthy target-language words for MT that reflect the contextual usage of the word. The merging effect for more than one word shows that the model can also be used in disambiguation tasks and as a lexical knowledge representation reference.

## References


Jean-Paul Benzécri. 1992. *Correspondence Analysis Handbook*. Marcel Dekker, New York.

Alexander Budanitsky and Graeme Hirst. 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, June.

Ido Dagan and Alon Itai. 1994. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4):563–596.

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

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

Philip Edmonds. 1997. Choosing the word most typical in context using a lexical co-occurrence network. In *Proceedings of the 35th Annual Meeting of the ACL*, pages 507–509.

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


[3]For word-sense disambiguation, clustering by cliques (not contexonyms) is a better choice, since it excludes shared cliques.


Hyungsuk Ji and Sabine Ploux. 2003. Automatic contexonym organizing model. In *Proceedings of the 25th annual meeting of the Cognitive Science Society*. In press.

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

Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse Processes*, 25:259–284.

Lillian Lee and Fernando Pereira. 1999. Distributional similarity models: Clustering vs. nearest neighbors. In *Proceedings of the 37th Annual Meeting of the ACL*, pages 33–40.

Dekang Lin and Patrick Pantel. 2002. Concept discovery from text. In *Proceedings of the*.

Christopher D. Manning. 1993. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st Annual Meeting of the ACL*, pages 235–242.

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

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

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

James Pustejovsky and Branimir Boguraev. 1994. Lexical knowledge representation and natural language processing. *Artificial Intelligence*, 63:193–223.

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

Eric Wehrli. 1998. Translating idioms. In *Proceedings of COLING98, Montreal*, pages 1388–1392.

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