The Semantic Atlas: an Interactive Model of Lexical Representation

Sabine Ploux*, Armelle Boussidan*, Hyungsuk Ji #

***** L2C2, Institut des Sciences Cognitives-CNRS, Université de Lyon, Bron, France

Intelligent HCI Convergence Research Center, Sungkyunkwan University, Seoul, Korea

sploux@isc.cnrs.fr; armelle.boussidan@isc.cnrs.fr; jihyungsuk@skku.edu

Abstract

In this paper we describe two geometrical models of meaning representation, the Semantic Atlas (SA) and the Automatic Contexonym Organizing Model (ACOM). The SA provides maps of meaning generated through correspondence factor analysis. The models can handle different types of word relations: synonymy in the SA and co-occurrence in ACOM. Their originality relies on an artifact called 'cliques' - a fine grained infra linguistic sub-unit of meaning. The SA is composed of several dictionaries and thesauri enhanced with a process of symmetrisation. It is currently available for French and English in monolingual versions as well as in a bilingual translation version. Other languages are under development and testing. ACOM deals with unannotated corpora. The models are used by research teams worldwide that investigate synonymy, translation processes, genre comparison, psycholinguistics and polysemy modeling. Both models can be consulted online via a flexible interface allowing for interactive navigation on http://dico.isc.cnrs.fr. This site is the most consulted address of the French National Center for Scientific Research's domain (CNRS), one of the major research bodies in France. The international interest it has triggered led us to initiate the process of going open source. In the meantime, all our databases are freely available on request.

1. The model

The originality of the model relies on the concept of 'cliques' (Ploux, 1997). Conceptually, a clique is a minimal unit of meaning, at a very fine grained level. In the SA cliques are lists of words which are related to each other by synonymy (in the broad sense). Mathematically, a clique is an object that designates a maximal, complete and connected subgraph. The graph here is a set of words (the nodes) and relations (the arcs) that link up words. Cliques differ from 'synsets' in Wordnet (Fellbaum, 1998), representations offered by LSA (Landauer *et al*., 2007), HAL (Burgess & Lund, 1997), or other word vector space models (Sahlgren, 2006). These models differ in their treatment of meaning, whether or not they use context relations, whether they use hierarchical links between words or proximity (distance within a vector model) and whether they are generated manually versus automatically and statistically. The SA offers a treatment of meaning at the level of relationships between words in the lexicon and ACOM at the level of co-occurrence in text (contexonyms). This treatment does not rely on a subjective hierarchical classification but on a geometrical model generated automatically and statistically which assigns vectors to cliques rather than words and therefore provides representation of word meaning based on units smaller than words. Unlike synsets, cliques are not units of language or metalanguage but rather an infra-linguistic tool. They allow for a very precise treatment of polysemy, which is the main aspect most word vector space models have trouble modeling since these models associate a single vector per word. Cliques provide a representation of meaning going beyond the lexical unit and thus also allow for sense navigation inside and outside the word boundary. In this respect cliques represent a conceptual level of meaning almost independent from words. They reflect the internal complexity and organization of the semantic structure of a word as well as the structural relationship of the word with others words (either within a lexicon, across lexicons or within a corpus).This makes the model a unique tool for translators who are well aware that words are not in a one to one correspondence. For a given word there is an underlying topology to the set of associated cliques that allows us to navigate from a semantic value to another in a continuous paradigm. Cliques organize meanings into value types, such as physical, emotional or perceptual aspects. Since each clique is connected to the next one by one synonym in common, a progressive transition from a meaning to another at subtle semantic levels is made possible. As an example, here are three pairs of cliques for the word *mind*, taken from the SA:

16: brain, head, mind, nous, psyche
17: brain, intellect, mentality, mind, wit
.
38: intention, mind, object, purpose, reason
39: intention, mind, spirit
.
43: mark, mind, note, notice, see
44: mark, mind, note, see, watch

On the basis of the list of cliques, a matrix composed of cliques as lines and words as columns is created. Then a semantic space is generated from the matrix. This method differs from other word space models that work with word/document matrices or word/paragraph matrices which soon encounter a problem of size limitation. Our model, however, has no size limits regarding corpora. With these matrices we calculate the coordinates of cliques with correspondence factor analysis (Benzécri, 1980). Unlike models that assign a vector or a node to a word, geometrical modeling associates a delimited space with a word in a continuous fashion. $\chi 2$ distance is used between words rather than Euclidean distance as it renders geometrical organization more accurately. Words are represented by the envelopes containing cliques. This way,

word envelopes may overlap or not, which provides a very instantaneous visual tool for definition or translation purposes. Finally a hierarchical classification algorithm generates clusters that organize senses. The number of clusters may be defined by the user to shed light on the specific aspects they wish to look at.

1.2 The model's parameters and graphical output

1 .2.1 SA parameters and output

The current online version of the model allows users to set a number of parameters at the query level and at the graphical output level. The model is set on default parameters to make all queries manageable at an optimal speed, but can be set to better serve specific querying. Online, up to five words at a time may be queried, using the + sign. This limit was set to better handle the numerous queries the site receives (up to 200.000 a day) but there is no limit as regards the number of words the model can handle as long as the output is decipherable. For a given search word the user can set a number of parameters before calculation and after calculation.

Query parameters (before calculation)
Users may:
-Set the search type for English words as standard (narrow synonymy) or enriched (broad synonymy). In French the default search uses broad synonymy since there is no distinction between narrow and broad synonymy in French lexicography.
- Choose the number of clusters they wish to visualize, calculated either on the basis of a number of cliques or words. Numerical options are optimal, maximal, or zero. The optimal setting adapts the output to the given input, while the maximal setting integrates all the results and the zero clique option only keeps the envelopes of words. This allows for highlighting one or many specific sub-meanings of a given word.
-Set the number of dimensions to be visualized (up to 15). The model will show core meanings in the first two dimensions; however minor meanings may be contained in the other dimensions. The former can be made to appear directly on the graphical interface by choosing the axes to bring to the fore.

Output parameters (after calculation)
The output is a flexible graphical map. Users can reorganize elements on the map in terms of axes, clusters, and the way lists of items are organized. The number of clusters is set to 3 by default but can be reset to any number via the 'class' option. 'Class 1' clusters elements on the basis of the center of gravity of words (envelopes) whereas 'class 2' clusters elements on the basis of clique sets.
Three levels of conceptual organization may be chosen from the list window: the clique, word or cluster. Items may also be organized in alphabetical order, cluster order or geometrical order (from closest to the center to peripheral).

1.2.2 ACOM's parameters and results

Many of the above parameters are valid for ACOM too. However some query options are specific to ACOM as regards co-occurrence settings.
The corpus on which the calculations are made may be chosen from a list of available corpora. Then one may set the number of most frequent words to be removed from the output (from 0 to 1000) and set the number of contexonyms shown.
Three other parameters allow for selecting the number of words or the percentage of words the calculation should be applied on. The first one (α) applies to search words, the second to their children (β) for second order co-occurrence and the third (γ) to cliques.
ACOM's results closely mirror human word association patterns as shown in Ji *et. al* (2008). The authors show that subject's responses in word association tasks matches ACOM's results well, therefore validating the model's cognitive relevance. On this task ACOM performed better than LSA.

2. Corpora and databases

All the databases listed below are accessible online and available on request. The model handles a synonym database, a translation database and a 'contexonym' (ACOM's co-occurrence) database.
The synonym database contains, for each entry word, a list of its synonyms, each indexed by the number of cliques it contains and a list of cliques that contain the entry word. The databases are complete for English and French and in progress for Spanish, Portuguese and Korean. The initial French database was made of seven synonym dictionaries which were compiled with a process of symmetrisation. [1] In English the database was extracted from several thesauri (including the Roget's thesaurus). To compile the databases the relationships between words were kept but not the thesauri's structure. The data was entirely restructured, merged and symmetrised automatically with the help of the model.
The translation databases contain for each entry word its equivalent in the target language and a list of cliques in both languages associated with it. The English-French, French-English as well as Spanish-French databases are complete, while Portuguese-French, Portuguese-English, Spanish-English, Korean-French, Korean-English, Korean-Spanish and Korean-Portuguese databases are in progress.
Synonym and translation databases are available for consultation online on http://dico.isc.cnrs.fr and http://dico.nv.isc.cnrs.fr.
The contexonym databases associate for each English or French token its frequency count from the corpus and a list of its sentence or paragraph co-occurrences each indexed with co-occurrence frequencies. At the moment the corpora available online are the BNC and Project

[1] These dictionaries are the *Bailly, Benac, Du Chazeaud, Guizot, Lafaye, Larousse* and *Robert*. Once compiled they provided 40 000 entries.

Gutenberg in English and five years of French newspapers' corpus (including "Le Monde"). However, other corpora may be dealt with by the model.

3. Examples

Several linguistic indicators come out in the representations produced by the SA and ACOM: Prototypicality and productivity may be evaluated at a glance by looking at the position of elements relative to the origin of axes, the number of cliques produced as well as the size, position and complexity of the space attributed to words.

Below, we illustrate the model with queries on a sample word –*mind*-, using an English synonym query, an English-French translation query and with ACOM, in that order. [2]

3.1 Monolingual Synonymy representations with the SA

The monolingual representation of *mind* when set to three clusters clearly shows one verbal and two nominal clusters that do not overlap but are linked up by weaker clusters (fig.1).

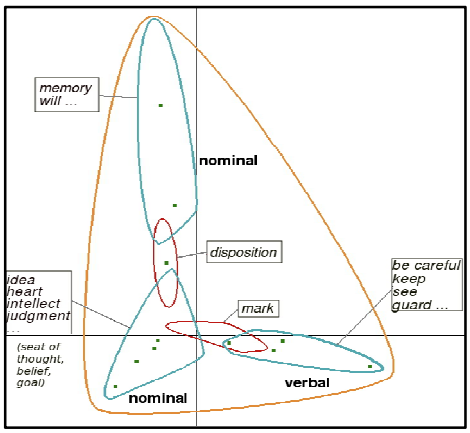

Figure 1: Illustration of a synonym query for *mind* with the SA. The main clusters are in blue, and the connecting clusters in red.

The meaning of *idea* is connected to the one of *will* via *disposition*, and *idea* is connected to *being careful* through *marking*. Meanings closer to the origin of axes are closer to a generic or core value determined by the model. Homonymous words will appear completely separate from this generic value. With cliques being organized in a continuous fashion, it is possible to navigate conceptually from verbal to nominal meanings and see which sub-meaning sets are shared or isolated. For instance the first verbal clique lists "1: *attend, attend to, look after, mind*" whereas no nominal clique retains the affective aspect found in '*to look after*'. However the following conceptual path : *to look after - to look - to watch - to see - to notice - to note - to mark - to bear in mind - to remember* - *memory* shows how one may navigate from a verbal value to a nominal one in a continuous fashion. This type of path may be explored graphically by moving the cursor over the map to make cliques appear, so that spatial organization may guide the understanding of semantic organization. The path may also be explored in more detail by looking at the cliques' list in the 'query information' link.

3.2 Translation (English-French) representations with the SA

The model can also be used in translation to distinguish values and determine which lexical unit is relevant (Ploux & Ji, 2003). The representation shows which senses are linked by a continuous conceptual path across the two languages and which ones are completely disconnected and do not overlap.

On fig.2 the envelope for *memory* is isolated and not connected to the others, as well as further away from the origin. This sense is the oldest etymological definition of the word *mind* and is obsolete (cf. Oxford English Dictionary). On the contrary, the two other sets show different sizes and relative strength in the two languages and a connecting concept between the two meaning areas.

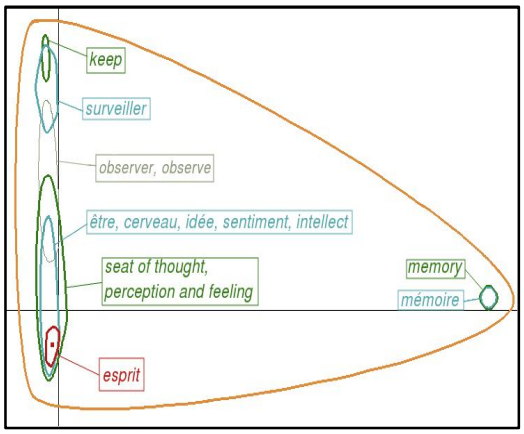

Figure 2: Illustration of a translation query for *mind* with the SA. English clusters are in blue, French ones in green. *Esprit*, in red, is the most probable candidate for translation.

The example of *mind* is particularly challenging since there is no straightforward translation of this word in the target language. The closest translation of *mind* is *esprit*; however *esprit* also means *spirit*. Where traditional MT systems might fail at choosing the best candidate, the graphical representation provides a tool to sort the possible candidates to translation. The map shows that *esprit* only represents a small space within *mind* and therefore is not a satisfactory translation in many cases. It is however very close to the origin, and therefore very close to *mind*'s core meaning (as determined by the model for this specific calculus). The number of cliques generated in each language is a first indicator of the extent to which lexical units match each other and how productive each sub-meaning is. For instance here we obtain 73 terms and 55 cliques for *mind* versus 104 terms and 156 cliques for *esprit*. This discrepancy highlights the number of senses

[2] The settings in the examples differ depending on what aspect of meaning is highlighted.

covered by *esprit* but not by *mind*. This partly explains why *esprit* may not be a good candidate since it carries a lot of meaning not included at all in *mind*. However both share the first most productive conceptual sub-group (alike to a semantic field) that may be defined as a set related to [intelligence], but differ at the level of the second most productive element, respectively [character] for *mind* and [belief] for *esprit*.

3.3 ACOM representations

The SA deals with synonymy but does not provide insight regarding word use in context. Paradigmatic relations may be sufficient in some types of research, but for a more complete approach syntagmatic relations are also necessary. Some words are weakly represented by synonymy, such as referential words (nouns of objects such as *notebook*) that may have a high frequency in a corpus but very few synonyms. In these cases, syntagmatic relations provide more information as to the word's semantics than paradigmatic ones. ACOM focuses mainly on syntagmatic relations as it exploits word co-occurrences in corpus. In includes some paradigmatic information as well since it incorporates second-order relations and uses broad co-occurrence windows. ACOM relies on cliques which are maximal sets of words in co-occurrence and selects words that co-occur in a homogeneous stylistic environment, giving information regarding genre, by default, whereas the SA provides a means to navigate the scale of genres independently of text. In this aspect the two models show complementary uses at the level of language and at the level of text.

In the following illustration (fig.3) we show how ACOM can be used to deal with the idiomatic aspect of language. The graph is not a default result but an example of a user's choice of visualization:

The collocations *frame of mind*, *peace of mind*, *mind (your) business* and *change (my) mind* are highlighted on top of the other meanings (*knowledge, thought*…). These collocations show and classify word use in context.

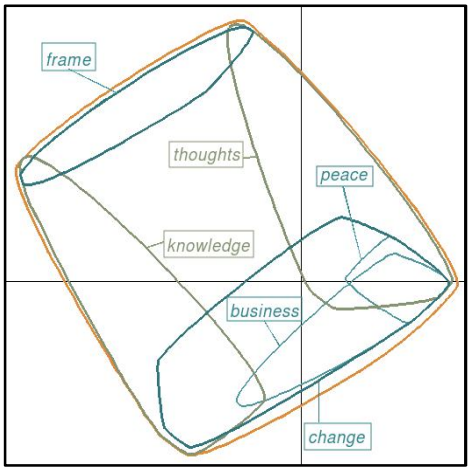

Figure 3: Illustration of an ACOM query for *mind*

4. Clusters

As illustrated in the above examples, clusters:

-Separate values within a monolingual representation on scales such as perception, action, emotion, etc.

-Detect overlapping and non-overlapping values across two languages.

-Separate values of a word's meaning according to its context of use.

Furthermore, as clusters allow for overlap detection, they may be used to analyse other types of linguistic data. For instance, clustering contributes to evaluating the impact of several factors over word meaning as shown in (Boussidan *et al.*, 2009a). In this case historical data - here etymological roots - was confronted with sound-form data - here phonaesthemes. Clustering is purposefully customizable so various linguistic needs may be met.

5. Current uses of the model and future perspectives

Currently, the model is used for the automatic treatment of polysemy and semantic disambiguation (Venant, 2007), and in two question answering projects relying on the synonym database to expand queries in order to cover the possible spectrum of senses more accurately (Grau *et al.*, 2006, and Project TCAN-CNRS 2003-2005). It has also been used in psycholinguistics research at the University of Geneva (Lachaud, 2005) in a survey to define the age of word acquisition in French as well as degrees of familiarity of words. Semi-automatic metaphor extraction and analysis may also be conducted with the model as shown in Oliveira & Ploux (to be published). Furthermore, various avenues of research are being explored to enrich and expand the model: Several languages are added to the translation databases. EEG studies in online word processing use the model's clustering tool for data analysis. Graphic research on meaning representation and structural properties of geometrical forms is conducted in a perspective of functional visualisation. Dynamic diachronic modelling to assess semantic change is in progress while specific phenomena such as neology is investigated as described in Boussidan *et al.* (2009b). The ACOM model also proved to be an excellent candidate for experimental cognitive tasks since it provides a good simulation of human responses as shown in Ji *et al.* (2008). Outside of the academic world, translators, journalists, writers and artists also use the model to guide their linguistic creativity.

We now aim at making the model open source, at easing its manageability as well as enhancing the availability of resources for all users to further profit the NLP, translation and psycholinguistics communities. We plan to make new corpus use more straightforward and to open the model to international multilingual collaborations in order to develop a collaborative semantic platform.

6. Conclusion

The AS and ACOM provide extremely fine-grained tools to detect and analyse linguistic patterns in corpora at the syntagmatic and paradigmatic level as well as across

languages. The models offer many possibilities in NLP, MT, psycholinguistics and cognitive science as well as outside the academic world. They are now evolving towards a more participative and dynamic format to allow several avenues of research to merge, and welcome collaborations in different sectors.

7. Acknowledgements

Thanks to Charlotte Franco, Masters student in Cognitive Science, and to all the people who participated in the model and databases (listed on http://dico.isc.cnrs.fr/en/membres.html)

8. References

Benzécri, J.-P. (1980). *L'analyse des données : l'analyse des correspondances.* Bordas, Paris.

Boussidan, A. Sagi, E. & Ploux, S. (2009a). Phonaesthemic and Etymological effects on the Distribution of Senses in Statistical Models of Semantics. *Proceedings of the Distributional Semantics beyond Concrete Concepts Workshop. 31th Annual Conference of the Cognitive Science Society.* Amsterdam, The Netherlands.

Boussidan, A. Lupone, S. Ploux, S. (2009b). La malbouffe : un cas de néologie et de glissement sémantique fulgurants. Atelier "Du thème au terme, émergence et lexicalisation des connaissances" *Actes de la 8ème conférence internationale Terminologie et Intelligence Artificielle*, Toulouse, France.

Burgess, C., & Lund, K. (1997). Modelling parsing constraints with high-dimensional context space. *Language and Cognitive Processes*, 12, 177-210.

Fellbaum, C. (Ed.) (1998). *Wordnet, an electronic lexical database*. Cambridge, Massachusetts: MIT Press.

Grau B., Ligozat A., Robba., Vilnat A. & Monceaux L. M. (2006). *FRASQUES: A Question-Answering System in the EQueR Evaluation Campaign* *LREC*, Genova, Italy.

Ji, H., Ploux, S. & Wehrli, E. (2003). Lexical Knowledge Representation with Contexonyms, in *Proceedings of the 9th MT summit*, 194-201.

Ji, H., Lemaire, B., Choo, H., Ploux, S. (2008). Testing the cognitive relevance of a geometric model on a word-association task: A comparison of humans, ACOM, and LSA. *Behavior Research Methods*, 40(4), 926-934.

Lachaud, C. (2005). La prégnance perceptive des mots parlés : une réponse au problème de la segmentation lexicale. Thèse de doctorat, Université de Genève.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2007). *Handbook of latent semantic analysis.* Mahwah, NJ: Erlbaum.

Oliveira, I. & Ploux, S. (to be published). Vers une méthode de détection et de traitement automatique de la métaphore. *Actes des journées scientifiques LTT 2009*, Lisbon, Portugal.

Ploux, S. (1997). Modélisation et traitement informatique de la synonymie. *Linguisticae Investigationes*, vol. 21, no. 1, pp. 1–28.

Ploux, S. & Ji, H. (2003). A Model for Matching Semantic Maps between Languages (French/English, English/French), *Computational Linguistics.* 29(2):155-178.

Sahlgren, M. (2006). The Word Space Model. Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. PhD Thesis. Stockholm University, Sweden.

Venant, F. (2007). Utiliser des classes de sélection distributionnelle pour désambiguïser les adjectifs, *TALN* , Toulouse, France.