

A Swedish Associative Thesaurus

*Person who says it cannot be done
Should not interrupt the person doing it.*
Chinese proverb (cit. after M. Hart)

In 1987 a computer-aided lexicographical project was initiated at Uppsala University Centre of Computational Linguistics. It was an attempt to organize the lexicon of a language in a coherent semantic network. The result is a full-scale dictionary, the "Swedish Associative Thesaurus" (*SAT*). It comprises more than 70 000 words, including, for example, all the entry words of the monolingual dictionary *Svensk ordbok* (1986). My co-workers on this project have been Gunilla Fredriksson and Agnes Kilar, the latter as programmer. The dictionary is unpublished: it exists in electronic form and in a few paper copies.

Originally, I was concerned with the problem of information retrieval and automatic indexation of texts. I imagined that one possible way to catch all meaning recurrence in texts would be to accumulate frequent meanings by transferring them not only from word forms to lemmas but also across lexical borders — to more central carriers of the very same meanings. It soon became clear that by successively connecting words and forming uninterrupted chains, we create a huge structure which can ultimately contain the whole lexicon of a language.

To my knowledge, *SAT* is quite a new type of dictionary. It shows words surrounded by their immediate semantic neighbours. The words occupy nodes in a tree-like (hierarchical) network; the arcs correspond to semantic relations between the words. We can regard these relations as resembling human kinship relations, which makes the network as a whole similar to a genealogical tree. Every word is accordingly surrounded by the members of its family – "parents", "children", "husbands"/"wives", and "siblings". The most important connection is that between a "child" and its "mother"; every word is assigned a descriptor called "mother". A "mother" of a word is its closest, more simple semantic neighbour. The criteria underlying these qualifications will be presented below. Thus, the word *dusty* is traced back to its "mother" *dust*, *book* to *read*, etc. If we assign *read* as a "mother" not only to *book* but also to *newspaper*, we create a "sibling" relation between *newspaper* and *book*.

In many cases it is desirable to connect a word with not only one, but two semantic neighbours, in order that the existing associations be more fully reflected. For instance, *gold* can be assigned the descriptors *metal* + *yellow*. The need for an extra descriptor is especially evident with regard to the compound words so typical of Swedish. Thus it is natural to trace a word such as *stenhus* 'stone house' back to a "mother" *hus* 'house' and a "father" *sten* 'stone'. The relation between *hus* and *sten* is thus one of a married couple: "wife" : "husband". Thus, every word in the lexicon is assigned an obligatory descriptor, a "mother", and an optional descriptor, a "father". The former descriptor often has a qualifying, the latter a modifying function, similar to the function of the second and the first part, respectively, of a compound word.

Certain principles have been elaborated to enable the choice of the descriptor called "mother". The concept of simplicity, or unmarkedness, is based on several criteria which, as a rule, support each other: morphological complexity, as in *dusty* : *dust*, frequency, as in *fair* :

beautiful, stylistic markedness, as in *guy* : *boy*, etc. Assymetrical semantic relationships also form a base for deciding which word is more simple. For instance, provided other criteria are met as well, hyperonyms are considered more simple than hyponyms (*copper* : *metal*), wholes more simple than parts (*roof* : *house*). In a case like *neigh* : *horse* it is important that *neigh* can hardly be thought of independently of *horse*, whereas *horse* enters into many relationships beyond that with *neigh*.

The concept of closeness is somewhat easier to handle than that of simplicity. It is based on: a) immediate semantic relationship; and b) determinism. For instance, *neigh* is a close neighbour of *horse*, whereas *oats* is not so close: horses eat other things than oats and oats are used for other purposes than feeding horses. In the same way the connection *trunk* : *elephant* is closer than that between *elephant* and *grey*: many things are grey, not only elephants, and colour is not a salient property of elephants.

When it is no longer possible to find a simpler neighbour in the vocabulary, we close the network at the top using the artificial word PRIM. This descriptor, which, by the way, can have no "spouse", has been given to about 50 words, such as *all*, *other*, *when*, *what*, *warm*, *can*, *must*, *know*. All these words function as descriptors of other words, although two of them, the extremely asemantic words *att* 'to' (before infinitives) / 'that' (conjunction) and *den* 'the' / 'that' (pronoun), give rise to just of a couple of grammatical terms.

We can think of the lexicon as located in a two- or three-dimensional space with one centre, PRIM, the position of every word being determined by its distance from the centre and its immediate semantic neighbours ("family"). The distance from the centre can be measured in terms of chains of obligatory descriptors ("mothers"). For instance, from *dammig* 'dusty' we can reach the centre through the following words: *damm* 'dust', *torr* 'dry', *våt* 'wet', *vatten* 'water', *ämne* 'substance', *vad* 'what', PRIM. (For the corresponding English words it would not necessarily be the same chain.)

The strict conditions regulating the choice of "mother" do not apply to the "father/husband". This word is not thoroughly checked with respect to simplicity, but it must surely be semantically related to the "child" — otherwise it would not do as a descriptor — and if there are several alternatives closeness is an important criterion. For instance, *kyrktupp* 'church weathercock' receives the "father" *kyrktorn* 'church tower' rather than the more distant descriptor *kyrka* 'church'. There may be quite a weak connection between "husband" and "wife". To take the same example, the words *kyrktorn* and *tupp* 'cock' would not be related in Swedish, were it not for the habit of placing a metal cock on the top of church towers. A "father" can hardly be a hyperonym of its "child", as its "mother" often is. Other relations, such as "part of", are possible, as in *kyrktorn*. In *altare* 'altar', which receives the single descriptor *kyrka* 'church', the same relation holds between "child" and "mother", since it is difficult to find any other more central descriptor in this case.

Needless to say, not only compounds may be provided with two descriptors. The Swedish word *näver*, meaning 'birch bark', is naturally given the descriptors *bark* + *björk* 'birch'. And vice versa, it is not always the case that a compound is analyzed according to its morphological structure; thus *järnväg* 'railway' receives only one descriptor, *tåg* 'train'. In cases where two descriptors are chosen independently of the morphological structure of the word it is not always possible to adhere to the principle that the first descriptor must be qualifying and the second modifying. Thus, *kyssa* 'kiss' is traced back to *läpp* 'lip' + *beröra* 'touch' (rather than the other way round), *läsa* 'read' to *se* 'see' + *veta* 'know', etc.

We can now also imagine further, more complicated, relations between "siblings": those which share a "mother" (*stenhus* : *trähus* 'wooden house') and those which share a "father" (*stenhus* : *stenyx* 'stone axe'). Unlike real life, even a more remote relation is possible: the "mother" of "sibling" A is the "father" of "sibling" B (*stenhus* : *hustomte* 'house gnome'). Words sharing both "parents" are closely connected. For example, the words *drälla*, *krylla*, *myllra*, *vimla*, all meaning approximately 'swarm', have the common descriptors *många* 'many' + *oordnad* 'unordered'.

The concept of "semantic relationship" is understood in a broad sense, including specific reference. Unlike traditional dictionaries SAT includes proper nouns. There are about 3 000 of them, for the most part words with a specific reference. They denote, for instance, politicians, writers, artists, countries, cities, rivers, mountains, organizations, institutions, and even novels and films. To take just one example (with 100 % determinability), the descriptors of *Paris* are: *capital* + *France*. As a rule, proper nouns receive common nouns as their "mother" descriptor.

The dictionary contains not only words, but also phrases. Phrases are, as a rule, traced back to one-word descriptors, e.g., *kick the bucket* would receive the descriptor *die*. The Swedish expression *på måfå* 'at random' has *slump*¹ 'chance' as its "mother" and *måfå* (a word not used outside this phrase) as its "child".

Words as morphological units are of no importance in this dictionary. If a word has several meanings, each meaning is treated as a separate lexical unit and is supplied with its own index number. This has the effect of making the number of indexed words very large.

The fact that "fathers" are selected with less rigidity than "mothers" results in a certain, justified and controlled, circularity. For instance, *segelbåt* 'sailing boat' is given the descriptors *båt* + *segel* 'sail, n'. But the "father" *sail*, in turn, has *boat* as its "mother". So *sail* is at the same time "child" and "husband" of *boat*. This is inevitable: sails are firmly connected with boats, but not every boat has sails. Another kind of permitted circularity is illustrated by *tallskog* 'pine forest', where the members of the "married couple" *tall* + *skog* are at the same time "siblings", since they have the same "mother", *tree*. However, the semantic relations holding between "mother" and "child" must in such cases be different.

A prohibited circularity could arise if we assigned descriptors automatically in pleonastic expressions like *isvak* 'ice-hole'. Here the compound is synonymous with, not a hyponym of, the second component. As with other synonyms, the more peripheral word is traced back to the more central one without any "father" being assigned. In this case we end up with the analysis *isvak* : *vak* : *hål* 'hole' + *is* 'ice'. On the other hand, "compressed" compounds like *tefat* (lit.) 'tea saucer', *kärnkraftverk* 'nuclear power plant' are given descriptors that reveal their real structure: *fat* + *tekopp* 'tea cup' and *kraftverk* + *kärnkraft*, respectively.

Large "sibling" groups have been avoided. Rather, they have a tendency to fall apart, thanks to a certain heterogeneity which makes it possible to differentiate their descriptions. However, purely morphological differences are not sufficient reason to split up semantically homogenous groups; for instance, compounds ending in *-lik*, *-liknande* are all assigned the same "father", *likna* 'be similar'.

Nevertheless, certain descriptors are used intensely. The following are used more than 200 times: *inte* 'not', *växt*¹ 'plant', *utan* 'without', *hon* 'she', *förnamn* 'first name', *djur* 'animal'. 74 descriptors have a number of "children" equal to or exceeding 70. Several of these descriptors serve as passages into special areas of the lexicon. Thus, a great many of the proper nouns are introduced by qualifying ("mother") descriptors like *förnamn*, *efternamn* 'surname', *författare* 'writer', *stad* 'city', *huvudstad* 'capital'. Large encyclopaedic areas, some with a rich

terminology, are introduced by qualifying descriptors like *djur*, *växt*, *vetenskap* 'science', *religion* 'religion', *musik* 'music', *språk* 'language', *sjukdom* 'disease', *maträtt* 'dish'. Important modifying ("father") descriptors, not only for encyclopaedic use, are *vatten* 'water', *fartyg* 'ship', *möjlig* 'possible', *före* 'before', *efter* 'after', *alltför* 'too'. One intensely used subgroup consists of negating descriptors such as the "fathers" *inte*, *utan*, *omöjlig* 'impossible' and the "mother" *avlägsna* 'remove'.

The material contained in this dictionary has been checked (by computer programs) to ensure that the formal conditions on the network are fulfilled everywhere: that no word lacks description or has been assigned two different descriptions; that there is no real, prohibited circularity; that indexed words everywhere have an index attached and that there are no gaps in the series of indexes; and so on.

To give all the information needed it is sufficient to store the material in one large, ordered or unordered, file containing all the words, supplied with their descriptions:

dusty : dust	trunk : nose + elephant	
dust : dry	dry : wet + not	
book : read	Paris : capital + France	etc.

But clearly this is an inconvenient way of presenting the dictionary to its users. I have therefore chosen to present the words in alphabetically ordered entries. In the original version every word constituted an entry of its own. Each entry could contain up to four sections, apart from the keyword: "parent(s)", "child(ren)", "spouse", and "sibling(s)". But a more efficient and less paper-consuming way (if the dictionary is printed out) is to show just the first two of these sections and integrate the third section into the second. Still another possibility is to let only words with "children" form entries; "childless" words are then to be found in the entries of their "parents". This reduces the number of entries from 72 000 to some 25 000.

Thus in the shorter version each entry contains three fields: a) the keyword itself, b) the "parent(s)", and c) the "child(ren)". The sections are indicated in the following way: 11 = "mother", 12 = "father", 21 = "child" with the keyword as a "mother", 22 "child" with the keyword as a "father". In section 21 the "father", if any, is indicated in parentheses preceding the "child(ren)". In section 22 the "mother", which is obligatory, is indicated in the same way. Let me give a couple of examples (the first one is shown only in part):

Example 1:

damm²: 11 torr¹; 21 damma¹, dammig, dammtuss, stoft, (kisel:) kiseldamm; (sten:) stendamm; 22 (avlägsna:) damma², dammsuga, dammtorka, (korn³:) dammkorn [...]

In a verbalized form this information can be rendered as follows:

*damm*² 'dust' (to be distinguished from *damm*¹ 'pond') has one descriptor, *torr*¹ 'dry'; as a "single mother", it has the "children" *damma*¹ 'raise dust', *dammig* 'dusty', *dammtuss* 'dust wad', *stoft* 'dust'; together with the "husband" *kisel* 'silicon' it has the "child" *kiseldamm* 'silicon dust'; together with the "husband" *sten* 'stone' it has the "child" *stendamm* 'stone dust'; together with the "wife" *avlägsna* 'remove' it has the "children" *damma*² 'wipe', *dammsuga* 'vacuum', *dammtorka* 'wipe with a cloth'; together with the "wife" *korn*³ 'grain, speck' it has the "child" *dammkorn* 'speck of dust' [...].

Not all compounds with *damm*² are to be found in this entry. On a deeper level we find, for instance, *dammråtta* lit. 'dust rat' with *dammtuss* as "mother" and *golv* 'floor' as "father". "Underneath" *dammsuga* we find *dammsugare* 'vacuum cleaner', etc.

Example 2 (with English translations added):

torr¹ 'dry': 11 våt 'wet'; 12 inte 'not'; 21 damm² 'dust', torka¹ 'wipe, let dry', torka² 'get dry', torrlägga 'make dry', (alldeles 'totally':) kruttorr 'dry as gunpowder', snustorr 'dry as snuff, (klimat 'climate':) arid, (skinn 'skin':) skinntorr 'scraggy', (sko¹ 'shoe'): torrskodd 'dry-shod', (väder 'weather':) torra³ 'dry spell', (öde¹ 'desert':) öken 'desert'; 22 (barrträd 'conifer':) torraka 'dead conifer', (slättmark 'level ground':) mo¹ 'sandy heath', (sprit 'alcohol':) torrsprit 'solid alcohol', (substans 'matter':) torrsustans 'dry matter'.

Having given a short description of *SAT*, I will now compare it with some other kinds of monolingual dictionaries. It might seem that the descriptors assigned to every word in the thesaurus are much like the definitions in a traditional, alphabetically ordered dictionary, only more restricted and deprived of syntax. However, there is one more important difference: in a traditional dictionary only the definitions of peripheral words are given in terms of more central (frequent, etc.) words. For the central part of the lexicon it often happens that the words contained in the definition are more peripheral, and consequently, less well-known to the user, than the word being defined. We may ask whether the definitions of words like *nose*, *sing* and the like really help the user in any way. In *SAT* there is no such break between the peripheral and the central region of the lexicon space.

Thanks to the rigid control imposed there are no such things in *SAT* as words used somewhere in a definition without being assigned a definition of their own. It is also impossible for a word to be indexed as an entry word but unindexed when used in a definition. Both these phenomena are quite common in traditional dictionaries. For example, in *Svensk Ordbok* the noun *sebu* is defined as *puckeloxe* 'hump-ox', but the latter word has no entry of its own; the adjective *slavisk* is split up in two indexed homonyms: *slavisk*¹ 'slavish' and *slavisk*² 'Slavic'. This adjective is used without index to define the (monosemous) verb *slavisera*: "göra (mera) slavisk" 'make more slavish/Slavic' (?); at this point we are left without help.

There is also a considerable difference between *SAT* and traditional thesauruses like Roget's classic *Thesaurus of English Words and Phrases* (originally from 1852) or *A Modern Thesaurus of English Synonyms* (1958). Here the words are distributed over certain main themes which are meant to reflect the structure of reality. I think this is not a very objective way of organizing the lexicon. It is true that *SAT*, too, is based upon a very large number of decisions made by a few individuals, but firstly, these decisions all have a very local effect, and secondly, there are ways of correcting these choices, the most important being the constant endeavour to create homogenous, even beautiful, "sibling" groups. So even if *prima facie* there are often several solutions, it turns out to be possible to decide which one is best. This is an effect of the pressure from the whole mass of surrounding lexical units. The result is, as I see it, a dictionary that does not impose a semantic grid upon the lexicon but reflects its inherent structure.

Another drawback with the traditional thesauruses is that they must be divided into two different parts, an index and the main dictionary. *SAT* has only one entrance, at least if we choose to let every word have its own entry.

A somewhat different way of creating a dictionary with, in principle, the same structure as a traditional thesaurus is my own *Rysk-Svenskt Minilexikon* 'Russian-Swedish Mini-Dictionary' (1993), which contains the upper frequency layer, about 4 000 words, of the Russian vocabulary. I started from an alphabetically ordered file and tried to bring together words which I considered semantically related. The result was a large number of "thickenings", which were eventually grouped together into 70 semantic regions, each one represented by a keyword. The ordering of the regions was then based on the part of speech of the keywords.

I shall now compare *SAT* with two Russian dictionaries. The first one is *Slovoobrazovatel'nyj slovar' russkogo jazyka* 'Russian Derivational Dictionary' (1985), in which each entry constitutes a derivational nest, i.e., one underived word with all its direct and indirect derivatives. The structure is somewhat similar to that of *SAT*, but it relies exclusively on the inner form (morphological structure) of the words. There are no connections between words containing different roots.

It must be admitted that *SAT*, too, in many cases relies on inner forms. For instance, the word *dammsugare* 'vacuum cleaner' is firmly associated with *damm* 'dust' in Swedish, but this connection is probably weaker in English. Clearly, the network depicted in this type of dictionary is in no way universal, but language specific.

The second Russian dictionary with which I wish to compare *SAT* is *Russkij asociativnyj slovar'* 'Russian Associative Dictionary' (1994). This dictionary is the result of a large-scale test in which some 600 Russian students participated. The informants were presented with one stimulus word and asked to react — immediately — to this word with another Russian word or phrase. Of course, in a test like this it is impossible to investigate the whole vocabulary. In this particular case, 1277 words were given as stimuli. The material is presented in two volumes, the first containing alphabetically ordered stimuli and — in frequency order — all the reactions given to each stimulus, the second presenting the material in reversed order, that is alphabetically ordered reactions and for each reaction all the stimuli — again in frequency order — producing this reaction. Needless to say, the number of reactions were much greater than the stimuli. For the sake of comparison with *SAT*, we shall look only at the most frequent reaction to certain stimuli. Here are a few examples with nouns: *vojna* 'war' — *mir* 'peace', *golova* 'head' — *bolit* 'aches', *drug* 'friend' — *vernyj* 'true', *student* 'student' — *bednyj* 'poor'.

In a way, the data contained in this Russian dictionary seem more reliable than those of *SAT*, since they are based on a large number of informants. Is it possible to use the data from *Russkij asociativnyj slovar'* in order to corroborate the decisions taken in *SAT*? As it turns out, this is almost impossible. Firstly, the reactions given in the test do not conform to the principles underlying the *SAT* network. The reaction can be more central (*orel* 'eagle' — *ptica* 'bird') as well as more peripheral (*metall* 'metal' — *železo* 'iron') than the stimulus, and there are also cases of circularity (*syn* 'son' — *doč'* 'daughter'; *doč'* — *syn*). Secondly, the predominant relationship between stimulus and reaction is syntagmatic (*postel'* 'bed' — *mjagkaja* 'soft', *sol'* 'salt' — *zemli* 'of earth'), which makes the reactions not very suitable as descriptors. Besides, in many cases there is quite a low degree of determinateness; for instance, quite typically, the most frequent reaction to *stena* 'wall' is *belaja* 'white' (64 informants), whereas *doma* 'of a house' occupies only third place (24 informants). Obviously, this Russian dictionary depicts some kind of associative structure different from that represented in *SAT* and in traditional dictionary definitions.

References

- Allén, Sture et al. (1986). *Svensk ordbok*. Stockholm (Språkdata).
- Karaulov, Ju. N. et al. (1994). *Russkij asociativnyj slovar'*, vols I-II. Moskva.
- Lewis, Norman (1958). *A Modern Thesaurus of English Synonyms*. New York.
- Lönngren, Lennart (1993). *Rysk-svenskt minilexikon*. Lund (Studentlitteratur).
- Roget's Thesaurus of English Words and Phrases*, new edition prep. by B. Kirkpatrick (1987). London (Longman).
- Tichonov, A. N. (1985). *Slovoobrazovatel'nyj slovar' russkogo jazyka*. Moskva.