Revisiting the Potentialities of a Mechanical Thesaurus

Uta Priss, L. John Old

Napier University, School of Computing,
u.priss@napier.ac.uk, j.old@napier.ac.uk

Abstract. This paper revisits the lattice-based thesaurus models which Margaret Masterman used for machine translation in the 1950's and 60's. Masterman's notions are mapped onto modern, Formal Concept Analysis (FCA) terminology and three of her thesaurus algorithms are formalised with FCA methods. The impact of the historical and social situatedness of Roget's Thesaurus on such algorithms is considered. The paper concludes by discussing connections between Masterman's research and modern research.

1 Introduction

In the 1950's and 60's Margaret Masterman, a pioneer in natural language processing and AI, conducted research using a thesaurus for machine translation. Her paper "Potentialities of a Mechanical Thesaurus" describes a lattice-based model of Roget's Thesaurus (Masterman, 1956). Since lattice-based models of Roget's Thesaurus have also been described in recent Formal Concept Analysis (FCA) research (Priss & Old, 2004, 2005, 2006, 2007, 2008), the idea for this paper is to revisit Masterman's ideas, to formalise her algorithms in FCA notation and to compare her ideas to modern research.

Until recently, it was quite difficult to access Masterman's research because many of her ideas were published in technical reports which were never widely distributed. Furthermore, her writing is influenced by the 1950's intellectual background, and is sometimes difficult to comprehend from a modern perspective. In 2005, however, an edition of her work was published by Wilks (Masterman, 2005) and supplemented with editorial comments and explanations which now make it possible to revisit Masterman's ideas.

Masterman was not a mathematician. Therefore her lattice-theoretical thesaurus notions are described mostly in prose and exemplified by descriptions of computer algorithms. Because the 1950's computers were quite different from modern technology, even the algorithms are sometimes obscure from a modern perspective. An example of the changes in computational technology is demonstrated by the fact that Masterman's group (the Cambridge Language Research Unit, CLRU) believed that "no computer could hold a coded thesaurus" (Masterman, 2005, p.105) because of the storage space required.

Instead of storing a coded thesaurus on a computer, the raw and uncoded thesaurus data was stored on punched cards. It is not clear, however, how much of the data was stored and in what format. Furthermore, it is not clear how much human invention was required in order to execute any of the thesaurus-based calculations. Perhaps the computer was only used to compare thesaurus categories, but the initial processing of the input data was done by hand? Masterman (2005, p.156) mentions that the "person operating the thesaurus must use his or her own judgement". Masterman's early experiments were only simulations of punched card programs because "the experiments could not be carried out automatically, as the CLRU had no computer" (ibid. p.159). But the lattice program, which was coded by A. F. Parker-Rhodes, "will in the near future actually go through a computer" (ibid. p.86). Wilks (1998) explains that Masterman later "had the thesaurus card punched (twice for error checking), which then formed the basis for a range of experiments ... performed on Hollerith sorting machines".

Another difficulty is the difference between editions of Roget's Thesaurus. If one had an exact copy of the thesaurus used by Masterman, one could approximate her algorithms by running them on the same data and comparing the results. Masterman mentions that "Roget's Thesaurus with additions was used" (Masterman et al., 1959) indicating that her thesaurus was amplified with further terms and edited for the purposes of her research. In some cases, the additions may have been ad hoc in response to a particular translation task (Masterman, 2005, p.153). Wilks (1998) comments that Masterman "had the whole of Roget's Thesaurus compacted from a thousand or more heads to eight hundred". He further states that Masterman's experiments "were not very successful because the heads were too sparse to give sufficient repetition." Thus, using a modern edition of Roget's Thesaurus is a disadvantage for the purpose of understanding Masterman's work, but it is actually an advantage with respect to examining the general validity of her work because the modern editions contain more words.

The next section[1] provides an overview of Masterman's work of using thesauri in machine translation. Section 3 describes Masterman's thesaurus model. Section 4 provides an FCA formalisation of three of her thesaurus algorithms. Section 5 discusses the changes between different editions of Roget's Thesaurus. Section 6 reconstructs one of Masterman's examples using a modern edition of the thesaurus and modern software. Section 7 traces the connections between Masterman's work and modern research.

2 Lattices and thesauri in mechanical translation

In 1956, at the International Conference on Mechanical Translation at MIT, four researchers from the CLRU (Masterman (1956), Richens (1956), Parker-Rhodes (1956), Halliday (1956)) reported on their research of using a thesaurus as an interlingua in "mechanical translation" (MT), the term then used for "machine translation". The group's founder, Masterman, envisioned using mathematical lattice theory for building a thesaurus, i.e. a hierarchical structure with groupings of synonyms or near synonyms. She thought that a "multilingual MT dictionary is analogous, in various respects, to a thesaurus" and that "the entries form, not trees, but algebraic lattices, with translation points at the meets of the sublattices" (Masterman, 1956). The advantage of this approach is that instead of having to consider different pairs of languages separately, each language needs to be translated only once (into the thesaurus). Adding a new language, then, does not require any changes to the previously added languages. Master-

[1] A draft version of Sections 2 and 7 is contained in: Priss & Old (2008). "Lattice-based Modelling of Thesauri." Lattice-Based Modeling Workshop, Palacky University, Czech Republic.

man stated that "the complexity of the entries need not increase greatly with the number of languages, since translation points can, and do, fall on one another" (ibid. p.36).

Of course, computational research in the 50s and 60s was influenced by the limitations of computers at that time. Considerations about computational speed and storage problems determined the algorithms. Parker-Rhodes (1956) extended Masterman's ideas by describing a mechanical translation program for interlingual thesauri using Boolean operations that "can be performed with very great speed". The storage problem would be solved by storing "all the relevant information ... in the input and output dictionaries". Richens (1956) described the algebraic interlingua, NUDE, its code and an overview of its translation operations.

MT algorithms at that time often started with a chunk-by-chunk literal translation (Masterman et al., 1959). Every word stem and every grammatical indicator was translated from the input language to the output language using a dictionary and some rules. This was also referred to as "pidgin translation" (Masterman, 2005, p.161). Masterman's use of lattices was novel because other linguists at that time (for example Lehmann (1978)) saw translation as a mapping between trees. A sentence from the input language was parsed into a tree structure. Each branch of the input language was mapped onto a branch of the output language. The branches in the output language formed another tree which had the output sentence as their root. Masterman argued that from a semantic viewpoint, lattices are a better model than trees. In a lattice, pairs of elements can have different numbers of parents and children, instead of having only one parent each in a tree structure. Thus, combinations of meanings can be represented more naturally. In 1965, Wilks commented that word-for-word pidgin translation was too limited because phrases, not words, are the semantic units of a sentence (Masterman, 2005, p.186). But this does not contradict the idea of using a thesaurus as a means of translating the semantic units.

In particular, Masterman (1957) was interested in Roget's Thesaurus (RT). Her idea was that each of the 1000 categories in RT could be used as a "head" which described the core meaning of a word. Because words that have more than one sense occur more than once in RT, a word can have several heads. This leads naturally to a lattice, not tree structure. Of course, this implies that the meets and joins need to be calculated; without meets and joins, a thesaurus would be just a partially ordered set, not a lattice. Multiple occurrences of a word in the thesaurus might correspond to different meanings of the word or even homographs (such as "lead" the verb and "lead" the metal). If one determines the heads of all the words of a sentence, the heads provide an indication of what the sentence is about. Individual words can be disambiguated by comparing their heads to the other heads in the sentence. If a word has two different heads and only one of these also occurs for other words in the same sentence, then it is quite likely that that head corresponds to the meaning of the word in this sentence.

Masterman et al. (1959) saw a relationship between MT and information retrieval because in both cases a thesaurus could be used: either for retrieval, or as an interlingua. Even grammar and syntax were dealt with by the thesaurus (Masterman, 1957) because grammatical indicators in the "intralinguistic context" relate to structures in the "extralinguistic context" that are shared across languages. For example, some languages have no genders (English), others have two (French), three (German) or six (Icelandic).

But the distinction between "male" and "female" is extralinguistically motivated (ibid. p.39). Masterman et al. (1959) see an interlingua as consisting of a "logical system giving the structural principle on which all languages are based". In modern terminology, the thesaurus represents the "conceptual structures" that underly information retrieval and natural languages. Because different languages share conceptual structures, they could share a thesaurus as a representation of conceptual structures. Thus, Masterman's ideas anticipated modern research into semantics, AI, knowledge representation and formal ontologies. The next section shows how her thesaurus notions can be mapped to their modern equivalents.

3 Masterman's thesaurus model

Masterman describes a thesaurus in terms of *heads*, *fans* and *tags*. Fig. 1 summarises her notions. Heads correspond to *contexts* (Masterman, 2005, p.109) or *situations* (ibid. p.194). Different situations are distinguished from each other by *similarity* and *contrast* (idid. p.189). In Roget's Thesaurus (RT), heads correspond to the approximately 1000 categories. The notion of *fan* refers to the polysemy of a word. A fan consists of a word and its different uses or meanings which could be changing over time (ibid. p.39). In RT, the index at the back of the book (or what Masterman et al. (1959, p.925) call the *cross-reference dictionary*) shows a fan for each word. Heads are further subdivided into *lists*, which are mutually exclusive, such as "spade, hoe, rake"; and *rows* which are quasi-synonymous, but non-exclusive, such as "coward, faint-heart, poltroon" (Masterman, 2005, p.109). Extralinguistically, the elements of a row can denote the same object, whereas the elements of a list denote different objects. Thus, the elements of a list "are sub-species of a genus, and not synonyms at all" (Wilks, 1998) or might form an extension of a natural language concept. In some editions of RT, lists and rows are visually differentiated because lists are printed in a one-word-per-line format. Rows are the smallest groupings of the words in RT: each category is subdivided into paragraphs which contain semicolon-delimited sets of words.

Masterman's notion	in Roget's Thesaurus	general notion
head (context/situation)	category	concept
similarity/contrast	(implicit)	attributes
list	list	extension
row	paragraph or semicolon group	synonymy
fan	index	polysemy
tag (aspect)	part of speech etc	facet
archeheads (aspect)	antonymy (in table of contents)	facet/frame/role

Fig. 1. Masterman's main thesaurus notions

Archeheads or *aspects* are additional classifications that crosscut the classification by the heads. Aspects are also distinguished from each other by similarity and contrast in the same manner as situations (ibid. p.189). Neither situations, nor aspects are permanently fixed, but instead both change over time and if different principles are applied (ibid. p.189). Masterman uses the metaphor of "re-sorting a pack of cards" in order to illustrate the flexibility of classifications (ibid. p.190). In older versions of RT, the main archeheads are "pleasing" and "non-pleasing" (ibid. p.110), because the categories alternate between a positive head followed by its antonymous counterpart. This is predominantly displayed in the table of contents of older versions of RT. In modern information science terminology, archeheads are *facets*. Masterman further uses the notion *(semantic) tag* to refer to aspects (ibid. p.202). In this case, this seems to include part of speech tags, which is a facet that is applied to all categories in RT. The other examples of tags given by Masterman (ibid. p.202 and p.204) appear to be similar to modern thematic roles, frames or case relations. These do not systematically occur in any of the printed editions of RT, but must have been manually constructed by Masterman's group.

In summary, Masterman's notions correspond to modern notions and closely relate to the structures that can be observed in Roget's Thesaurus. Her notion of *double classification* (ibid. p.190) using situations/heads and aspects/tags is similar to the modern notion of faceted classification. Her notions of heads, similarity/contrast and lists can be mapped onto FCA terminology. Heads, words and fans provide the basic formal contexts for an FCA modelling of thesauri as shown in the next section.

4 Three algorithms

This section describes three of Masterman's thesaurus algorithms in FCA terminology. All three algorithms follow the same core structure and are similar to what are called *restricted neighbourhood lattices* (Priss & Old, 2004) in modern research. A (modern) neighbourhood context in Roget's Thesaurus starts with a word, looks up all the senses (rows or heads) of this word, and then looks up all the other words that share the same senses. This process can be started with either words or senses and is stopped after several iterations. The usual FCA operator for finding all attributes that belong to a set of objects is the *prime operator*[2]. The *plus operator*[3] used for constructing neighbourhood lattices differs from the prime operator in that it looks for attributes belonging to *any* (instead of *all*) of the objects. Thus, it usually enlarges the sets. This enlargement can be reduced by requiring that attributes need to be shared by at least two (or any other number) of objects[4]. Such neighbourhood lattices are called *restricted* (Priss & Old, 2004). Applied once to a single object (or single attribute), the prime and plus operators yield the same result ($g' := G_1' = G_1^+ =: g^+$ for $G_1 = \{g\}$).

The first algorithm considered here is a translation of a Latin sentence "Agricola in curvo terram dimovit aratro" into English[5] (Masterman et al. (1957) and Masterman (2005, p. 149)). It involves the following stages:

[2] For a formal context $(G, M, I), G_1 \subseteq G, M_1 \subseteq M$, $G_1' := \{m \in M \mid gIm$ for all $g \in G_1\}$. $M_1' := \{g \in G \mid gIm$ for all $m \in M_1\}$. To avoid ambiguity, G_1' can also be written as G_1^I.

[3] $G_1^+ := \{m \in M \mid gIm$ for one $g \in G_1\}$. $M_1^+ := \{g \in G \mid gIm$ for one $m \in M_1\}$.

[4] $G_1^{(I,\geq n)} := \{m \in M \mid gIm$ for at least n elements $g \in G_1\}$. Thus, $G_1^+ = G_1^{(I,\geq 1)}$.

[5] For example: "The farmer parts the earth with his curved plough."

1. dictionary matching: for each chunk of the input language a set of English heads is found representing semantic, syntactic and grammatical elements
2. operations on semantic heads: basic translation
3. operations on syntactic heads: syntactically complete, unparsed output
4. operations on grammatical heads: parsed and ordered output
5. cleaning up operations

Masterman argues that, in principle, grammar and syntax can be dealt with in the same manner (i.e., at a conceptual level using a thesaurus) as semantics (Masterman, 2005, p201). But in Masterman et al. (1957), only the application of Roget's Thesaurus to Stage 2 is shown, which involves operations on semantic heads. Translated into an FCA notation, a thesaurus contains a formal context (W, H, F) of words $w \in W$, heads $h \in H$ and fans, which describe the relationship between a word and its heads. Masterman's algorithm starts by calculating $S^F = \bigcup_{w \in S} w^F$ which contains the heads w^F for each chunk w of the input sentence $S \subseteq W$. Each head that does not occur in at least one other set of heads is eliminated: $\mathcal{H}(S) := \bigcup_{w \in S} w^{F/S}$ with $w^{F/S} := \bigcup_{v \in S, w \neq v} v^F \cap w^F$. This step eliminates homographic or polysemous senses that do not relate to the main topic of the sentence. It follows that $w^{F/S} = w^F \cap \mathcal{H}(S)$ and $\mathcal{H}(S) = S^{(F,\geq 2)}$. In modern terminology, Masterman's algorithm produces a restricted neighbourhood context $(S, \mathcal{H}(S), F \cap S \times \mathcal{H}(S))$.

If $w^F \cap v^F = \emptyset$ for all $v \in S, w \neq v$, none of the heads in w^F are in $\mathcal{H}(S)$. Thus, in order to determine a translation of w, the next higher grouping above the heads according to the table of contents of RT is considered. This can be represented as a formal context (H, A, C) of heads H, higher level classes A and the classification relationship C. Two plus operators are applied: first, the higher level classes are obtained, then all other heads that belong to these higher level classes are selected. This can be formally represented[6] as $w^{(F \circ C, \geq 1)(C, \geq 1)} = \{h \in H \mid \exists_{a \in A} \exists_{i \in H} : wFi, iCa, hCa\}$. This contains all heads that are in the same higher level classes as the original heads in w^F. Now restriction can again be applied: $w^{F/S*}$ is defined as $w^{F/S}$ if this is not $\emptyset$ and $\bigcup_{v \in S, w \neq v} v^F \cap w^{(F \circ C, \geq 1)(C, \geq 1)}$, otherwise. The result is a set $\mathcal{H}(S)^* := \bigcup_{w \in S} w^{F/S*}$ and a restricted neighbourhood context $(S, \mathcal{H}(S)^*, F \cap S \times \mathcal{H}(S)^*)$.

Each head in $\mathcal{H}(S)^*$ can now be translated into the target language. The target language is represented by a formal context (W_1, H, F_1) corresponding to the words W_1 and fans F_1 of the target language and the same heads as in the source language. Possible translations are elements of $\mathcal{H}(S)^{*F_1} \subseteq W_1$. More specifically, the translation of a word $w \in W$ with respect to S is provided by $T(w) := (w^{F/S*})^{(F_1, \geq 2)} = \{v \in W_1 \mid vF_1h$ for at least 2 elements $h \in w^{F/S*}\}$. Only those translations are selected that occur in at least two heads. The whole algorithm resembles the calculation of restricted neighbourhood lattices, with some modifications.

Some entries in Roget's Thesaurus contain cross-references. This is not to be confused with Masterman's "cross-reference dictionary" in the back of the book. Cross-references in the first half of the book are mostly a space-saving mechanism in RT: if a group of words occurs in two paragraphs, a cross-reference is used in one of the categories instead of duplicating the words in both places. In general, the use of

[6] $F \circ C$ denotes the relational composition. $w^{(F \circ C, \geq 1)} = \{a \in A \mid \exists_{i \in H} : wFi, iCa\}$.

cross-references is inconsistent in RT. Masterman argues that RT is a lattice where the joins are given by the hierarchy that is contained in the table of contents of the book, and the meets are given by the cross-references (Masterman, 2005, p.210). Thus, for the purposes of the translation algorithm, Masterman considers both words and cross-references that are shared between two heads. If a cross-reference is shared, it is added to its $w^{F/S*}$ and included in the computation of $T(w)$ as an iterative process. Masterman's reasoning for this is that "the numerical cross-references are to be interpreted ... as the overlap in meaning between the heads" (ibid. p.204).

In the Latin example sentence, this procedure leads to between one and eight possible translations for each word of the sentence (ibid. p.157). Similar results can be obtained with a modern thesaurus (see Section 6). Masterman explains that syntactic and grammatical operations would need to be applied in order to obtain the full translation, but these operations are not further explained. Incidentally, we ran the Latin sentence through a free online InterTran translation website[7] yielding the translation: "farmer upon to bow earth dimovit plowman". This is significantly worse than Masterman's head translations: "farm, farmer, bend/crook, ground/soil, plough/till, ploughman" (ibid. p.157). Unfortunately, we were not able to compare the result to Google's and Yahoo's translations (which yield good translation results in some cases) because these do not translate from Latin.

The second algorithm is very similar to the first but is applied to the translation of a whole passage of text instead of just a single sentence. It was originally described in Masterman (1956), but also appears in a slightly amended version in the appendix of Masterman et al. (1959). This algorithm consists of four stages:

1. Using a bilingual pidgin dictionary, chunks of the input language are translated into chunks of the output language.
2. Lattice position indicators are used for identifying the syntactic elements.
3. The thesaurus cross-reference dictionary is used to find the heads for the chunks.
4. The thesaurus procedure is applied.

Unfortunately, stage 2 is not explained in the paper. It is not clear how the syntax lattice is constructed. The main purpose of stage 2 appears to be to determine which heads to include in the calculation of $w^{F/S}$ (using the FCA notation from above). When dealing with a whole paragraph of text, it would not be useful to compare all heads simultaneously. Instead, only those heads are compared which are equal or subordinate to each other, based on the syntax lattice. As stated before, this is not completely clear in the paper, but it probably means that the syntax lattice is similar to a tree-parsed structure. Heads are compared if they are in a parent-child relationship in the tree. For example, the head of an adjective is compared with the head of its noun, and so on.

Otherwise, the algorithm is very similar to the previous ones, except that more procedures are added for dealing with cross-references and for adding words to the thesaurus. Most likely the reason for this is that the thesaurus used by Masterman was fairly sparse and did not yield any results for some of the head comparisons. Masterman argues that these procedures are not ad hoc, but instead make use of cross-references and of interpreting phrases and metaphors that are implicitly contained in the thesaurus.

[7] www.stars21.com and www.translation-guide.com.

She asserts that “all possible chains of meanings are somewhere in Roget’s Thesaurus if they can be found” (Masterman, 2005, p.93)[8]. The conclusion of Masterman’s paper is that the algorithm is able to improve the initial pidgin translation by replacing some of the words with more appropriate choices.

The third algorithm (Masterman, 1961) translates an English sentence into a thesaurus representation, which serves as an interlingual or conceptual representation. Even though we argue that Masterman’s algorithms resemble restricted neighbourhood lattices, she does not represent the results as lattices herself, though her 1961-paper contains a graphical representation of fans which are quite similar to our representations of neighbourhood lattices. *Jointed fans* (Masterman, 2005, p.207) are partially ordered sets that have a word at the top, and the different uses or meanings of the word below (as given in the index of Roget’s Thesaurus). In the diagram, each meaning is indicated by the number and the name of the head in which this word occurs. The resulting diagrams strongly resemble our representation of neighbourhood lattices (as in Figure 2). Jointed fans are trees, not lattices, but the smallest neighbourhood lattices (consisting of a word, its senses and other words that share these senses) tend to be mostly tree-like. Non-tree like structures emerge in neighbourhood lattices when the plus operator is applied more than twice, or when the algorithm starts with more than one word or sense.

The fans are further subdivided by the different parts of speech of the word under each head. The part of speech indicators are examples of *tags*. Masterman discusses the manual addition of further tags (ibid. p.204), which assign aspects to each occurrence of each word in the thesaurus. Although the words used for the tags are idiosyncratic, the resulting structure resembles a formal ontology. Masterman (ibid. p.209) argues that this fan structure, consisting of heads and tags, corresponds to dictionary entries and that, for translation purposes, dictionaries of different languages can be connected to an interlingual thesaurus using such fan structures.

The algorithm for translating the English sentence into a thesaurus representation given by Masterman (1961) is as follows (translated into FCA notation): for the thesaurus context (W, H, F) and a sentence S with words $w \in S$, the set $S^F = \bigcup_{w \in S} w^F$ contains the sets of heads for each word. For each element in $\mathcal{H}(S)$, the translation is provided as: $T(w) := (w^F)^{(F,\geq 2)} = \{v \in W \mid vFh \text{ for at least 2 elements } h \in w^F\}$. This is basically the same procedure as described above (Masterman et al., 1957), again resembling neighbourhood lattices. In this case, Masterman provides special rules for the instances where $T(w)$ contains more than one word or w^F contains zero or one head, but these rules rely on other non-thesaurus resources (such as special dictionaries).

[8] We believe that her assertion is correct. We have conducted experiments with changing the degree of restriction in neighbourhood lattices (Priss & Old, 2008) and found that it is possible to use heuristics to enlarge the lattices in cases where not much overlap is found. But, of course, the implementation of such techniques would have been difficult to imagine with 1950s computers.

5 Changes between the different thesaurus editions

If Masterman's experiments are to be replicated, it is of interest to determine how far they depend on the particular versions of the thesaurus. This section provides an excursus into the differences between the thesaurus editions in order to highlight the historical changes of the thesaurus categories.

There are two main versions of Roget's Thesaurus (RT) publications: the British line, published by Longman/Penguin from 1852-2002; and the US American line, branching from the British line in 1911, published by Crowell/Harper Row, and commonly referred to as the International version (RIT). Both contain the hierarchical classification system, the Synopsis of Categories, based on Roget's 1000 categories, with six to eight classes at the highest level. By default, the US thesaurus uses American spellings for common words such as color and jail, while the British version lists colour and gaol in the index. It is notable that the current International version does not list gaol, even as a Britishism, while the British version does list jail. This could reflect the influence of American spelling on British English. Masterman used RT (1953) for her experiments. Our own experiments are based on an amended version of RIT (1962)[9]. Because of copyright reasons, RT (1911) is the most current edition in the public domain.

With each new edition of RT, categories may be renamed, added, and sometimes combined or split (very, very rarely deleted). Among the reasons to change the name of a category, modernisation is the most usual. For example, Cicuration (RT, 1852-1953; RIT 1911-1922), meaning the act of taming animals. This was modernised to Animal Culture (RIT, 1946), then Animal Husbandry (RIT, 1962; RT, 1972). A further example is Preterition (RT 1852-1972; RIT 1911-1946), meaning passing, or passed. This was modernised to The Past (RIT, 1962) and Past Time (RT, 1972). Categories may be split where distinct ideas were obviously combined under one head (John L. Roget, 1933, Editors Preface); or expanded, for example, from "Earliness" (RIT, 1922) to "Earliness; Punctuality" (RIT, 1946).

Both versions of RT have in the past borrowed entries from the other, and occasionally, category names: Non-addition; Subduction (RT, 1852-1953; RIT, 1911-1922), Deduction (RIT,1946); Subduction (RT, 1962), Subtraction (RT, 1982; RIT 1992). Or perhaps these changes simply reflect the changing face of modern English (a term which is always current, no matter in which decade it is used).

The pattern of category changes is different from the British version to the American version. For example the updating of equivalent categories occurs earlier for the American editions. Also, the British version is more tolerant of obsolete (from daily usage) Latin terms. In recent years the addition of new words has mainly been in the areas of science and technology. The addition of these terms also differs between the versions. The British version continues to add new scientific and technical terms to existing Categories, while the American version has added new categories. For example, electronics is found under 160 Power in the British version (RT, 2002), along with strength, force and energy; while the American version has its own category for 342 Electronics (RIT, 1962).

[9] Which was converted to electronic form by Sedelow & Sedelow (1993) under NSF grants and later converted to a relational format by the second author.

The addition of categories can reflect cultural and political attitudes, just as Roget's original categories reflected the attitudes of his day. For example, category 986 Pseudo-revelation (original 1852 edition) contrasts the heathen Koran, Buddha, the Upanishads, and others, with orthodox Judeo-Christian beliefs of the time (category 985 Revelation).

The 1953 British edition was chosen by Masterman and the CLRU at a time when the paranoia of Macarthyism was rampant in the USA. Not coincidentally, much of their machine translation research was supported by the US military (National Science Foundation, U.S. Air Force Office of Scientific Research, and the Office of Naval Research (Washington, D.C), among others.

This had no apparent direct effect on RT, but had a massive influence on the ensuing US edition (RIT, 1962). That edition was used directly in the machine translation efforts of the US military to translate Russian military strategy. The following categories were some of those which were either added, or radically expanded: 277 Aeronautics, 280 Rockets and Flying Missiles, 281 Astronautics, 326 Radiation and Radioactivity, 342 Electronics, 345 Radar and Radiolocators, and 348 Automation. A sample of the type of words added is:

> 277 Aeronautics: aircraft hydraulics, jet engineering, kinetics, micrometry, rocket engineering, supersonics, supersonic aerodynamics; aviation medicine, Air Force School of Aviation Medicine, Air Force Department of Space Medicine; aerial navigation, celestial navigation, electronic navigation, automatic electronic navigation, navar, teleran, loran, shoran
>
> 348 Automation: robotic control, cybernetics, automatic electronic navigation, automatic guidance, missile guidance; guided missile, thinking machine, chess-playing machine; ENIAC UNIVAC, IBM 702

Some of the more cryptic words are acronyms, probably completely unknown to normal English speakers. For example, loran (Long Range Aid to Navigation) and teleran (Television Air Radar Navigation).

As this was also the beginning of the space race, many related words were added to categories such as 374 Universe (to do with astronomy, star systems, constellations, along with some navigational terminology). 279 Aircraft, in the 1946 (post-war) edition, had extensive lists of Second World War US, British, German, Italian and Japanese military aircraft. In 1962 these were replaced by terms such as "air-sea rescue amphibian, anti-submarine patrol; constant-chordrotor helicopter, intermeshing-rotor helicopter; high-altitude reconnaissance plane, long-range patrol bomber, photo-reconnaissance plane".

This excursus demonstrates the historical and social situatedness of thesaurus versions. Any thesaurus algorithm will be influenced to some degree by these versional changes. The actual labelling of the heads is not so important for neighbourhood lattices, but the types and numbers of words in the thesaurus and the degree of granularity in the lowest level groupings are relevant. Because the modern versions are copyrighted and not freely available in electronic format, it is difficult to conduct experiments that evaluate the impact of the versional changes. Our experiments have show that RT (1911) is too sparse compared to RIT (1962). Masterman's RT (1953) is similarly sparse as RT (1911).

6 Revisiting one of Masterman's examples

In this section, we revisit Masterman's treatment of the Latin sentence "Agricola in curvo terram dimovit aratro" as described in Section 4. We use Masterman's heads ($\mathcal{H}(S)^*$), but manually map them to the categories of RIT (1962) instead of RT (1953). In most cases, the category names are the same in both thesauri, only the numbers are different. For example, Agriculture is category 412 in RIT (1962) and category 371 in RT (1953). We then use our software as described in previous papers (Priss & Old 2004, 2007 and 2008) to calculate neighbourhood lattices for each $T(w)$. The six lattices in Figure 2 correspond to the translations of the six chunks "AGRI", "COL", "INCURV", "TERR", "DIMOV" and "AR". Most of the objects in these lattices are the words that Masterman retrieved as well. RIT (1962) retrieves slightly more words, than RT (1953).

This example confirms that even with the differences between thesaurus editions, the main structures are still similar. Because we started with Masterman's heads ($\mathcal{H}(S)^*$), this example only confirms the last part of Masterman's algorithm. With respect to the first part (the initial selection of heads), one would need a thesaurus of Latin words. Masterman conducted this part manually. The second part (the use of higher level classes and cross-references) is possible with RIT (1962), but so far we have not yet implemented it. Another confirmation of Masterman's research is the fact that her algorithms are similar to techniques that were independently developed by other researchers as discussed in the next section.

7 Modern descendants

Masterman's research influenced many people, including Karen Spärck Jones who is considered to be one of the pioneers in information retrieval and natural language processing. Spärck Jones used Roget's Thesaurus, but as far as we know had not much interest in lattices. Similarly, Yarowsky (1992) described an implementation of the use of Roget's for word-sense disambiguation which was very similar to Masterman's ideas (although he does not cite her), but he uses statistical methods instead of lattices.

In 1960s in the US, Sally Yeates Sedelow obtained funding to convert the American edition of Roget's (1962) into a machine readable format with the purpose of aiding machine translation. The initial abstract models that she and her husband, Walter Sedelow Jr., used did not rely on lattice theory (Dillon (1971), Bryan (1973), Bryan (1974), Talburt & Mooney (1989)). But Bryan's model describes a binary relation between words and senses which is very similar to a formal context as used in FCA. Thus, when the Sedelows met Rudolf Wille, the founder of FCA, in the early 1990s, they were enthusiastic about the possibilities that lattice theory had to offer for their research. Their paper about the concept "concept" (Sedelow & Sedelow, 1993) derives semantic neighbourhoods for words from the thesaurus which are then represented as "neighbourhood lattices". Our own research has used and elaborated this technique in a variety of papers (Priss & Old, 2004) and has recently led to the implementation of an on-line interface[10] that allows users to interactively generate such lattices. As explained above, Masterman's thesaurus algorithms are a form of "restricted neighbourhood lattices". Thus,

[10] `http://www.roget.org`

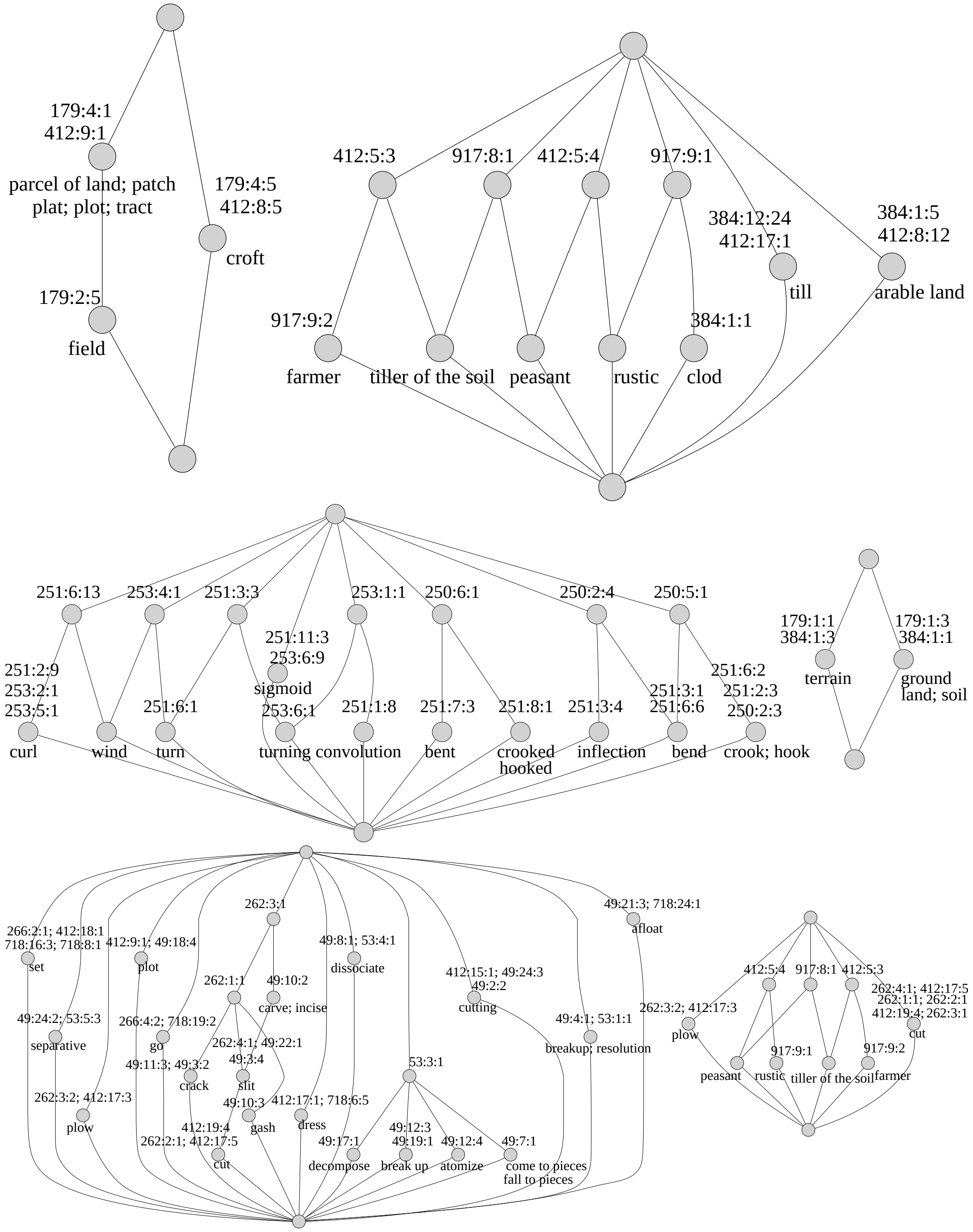

Fig. 2. Neighbourhood lattices for the translation of “Agricola in curvo terram dimovit aratro”

one can argue that this modern research is an implementation of Masterman's ideas, although the thesaurus research (of the Sedelows) was initially separated from lattice research and was only recombined through FCA.

Another modern instantiation of Masterman's ideas is Helge Dyvik's (2004) research, although, as far as we know, he was not directly influenced by, or aware of, either FCA or Masterman. Dyvik's lattices are feature lattices in the sense of componential semantics. Dyvik's Semantic Mirrors Method extracts semantic information from bilingual corpora. His assumption is that if the same sentence is expressed in two different languages, then it should be possible to align words or phrases in one language with the corresponding words or phrases in the other language using statistical processes or semi-automated processes. Once the corpora are aligned the "translational images" of words in the other language are computed. This process can be repeated several times. Next, the translational images are algorithmically assigned to separate senses. The resulting structures can be represented either graphically as lattices, or as a thesaurus (using a WordNet-style representation). Both structures can be generated interactively through an on-line interface[11]. Priss & Old (2005) have shown that this procedure is similar to creating neighbourhood lattices in FCA, though Dyvik's research was developed independently of FCA.

It could be argued that Dyvik's Semantic Mirrors method is a proof of concept for Masterman's vision. Masterman's (1956) statement that a "multilingual MT dictionary is analogous in various respects, to a thesaurus" and that "the entries form, not trees, but algebraic lattices, with translation points at the meets of the sublattices" prescribes exactly what Dyvik has implemented. Of course, it would not have been possible to implement a system like Dyvik's in the 1950s or 60s due to the limits of computers at that time. It seems to us, however, that perhaps not all of Masterman's ideas have fully been explored using modern technology. For example, the "Twenty questions method of analysis" (Masterman et al., 1959) that was used for extracting extralinguistic (or "semantic") information via an intralingual analysis, appears to be similar to attribute exploration in FCA (Ganter & Wille, 1999). But this relationship has not yet been further investigated.

8 Conclusion

In summary, Masterman's theoretical notions can easily be mapped to modern FCA notions. Her lattice-thesaurus research has been independently rediscovered (Dyvik (2004) and Sedelow & Sedelow (1993)) and has been implemented in recent FCA research (Priss & Old, 2004, 2005, 2006, 2007, 2008). The core of the algorithms and graphical representations described by Masterman resembles restricted neighbourhood lattices, which have been shown to be the most useful lattices for Roget's Thesaurus (Priss & Old, 2008). Several of her ideas are novel with respect to neighbourhood lattices of Roget's Thesaurus and could inspire future research:

- The use of cross-references. In library thesauri, "related term" links are similar to cross-references in Roget's Thesaurus. Thus, Masterman's approach for using cross-references is relevant for such thesauri.

[11] http://ling.uib.no/helge/mirrwebguide.html

- The use of an additional aspect classification (or faceted classification). Masterman suggests to modify the hierarchy of heads that is represented in the table of contents of Roget's Thesaurus so that it becomes a lattice structure in order to obtain a more detailed semantic representation. This can be achieved by adding semantic tags, and would be quite easy to implement for modern thesauri.
- Semantic tags for the individual occurrences of the words in the thesaurus. Masterman explains this procedure with a manually constructed example for one head. Clearly, it would be difficult to implement this for the whole thesaurus. But perhaps there are modern natural language processing algorithms that could be used for such purposes. At least the tags that already occur in the thesaurus (mostly part of speech tags) could be used in the process of constructing neighbourhood lattices.

References

1. Bryan, Robert (1973). *Abstract Thesauri and Graph Theory Applications to Thesaurus Research*. In: Sedelow, Sally Y. (ed.). Automated Language Analysis. Report on research 1972-73, University of Kansas, Lawrence.
2. Bryan, Robert (1974). *Modeling in Thesaurus Research*. In: Sedelow, Sally Y. (ed.). Automated Language Analysis. Report on research 1973-74, University of Kansas, Lawrence.
3. Dillon, Martin; Wagner, David J. (1971). *Models of thesauri and their applications.* In: Sedelow, Sally Y. (ed.). Automated Analysis of Language Style and Structure in Technical and other Documents. Technical Report, University of Kansas, Lawrence.
4. Dyvik, H. (2004). *Translations as semantic mirrors: from parallel corpus to wordnet.* Language and Computers, 49, 1, Rodopi, p. 311-326.
5. Ganter, Bernhard, & Wille, Rudolf (1999). *Formal Concept Analysis. Mathematical Foundations.* Berlin-Heidelberg-New York: Springer.
6. Lehmann, Winfred P. ; Pflueger, Solveig M. ; Hewitt, Helen-Jo J. ; Amsler, Robert A. ; Smith, Howard R. (1978). *Linguistic Documentation of Metal System.* Final technical report, Rome Air Development Center, RADC-TR-78-100.
7. Masterman, Margaret (1956). *Potentialities of a Mechanical Thesaurus.* MIT Conference on Mechanical Translation, CLRU Typescript. [Abstract]. In: Report on research: Cambridge Language Research Unit. Mechanical Translation 3, 2, p. 36. Full paper in: Masterman (2005).
8. Masterman, Margaret (1957). *The Thesaurus in Syntax and Semantics.* Mechanical Translation, 4, 1 and 2, p. 35-43.
9. Masterman, Margaret; Needham Roger M.; Spärck-Jones, Karen; Mayoh, B. (1957). *Agricola in curvo terram dimovit aratro.* CLRU memo ML-84. In: Masterman (2005).
10. Masterman, Margaret; Needham Roger M.; Sparck-Jones, Karen (1959). *The Analogy between Mechanical Translation and Library Retrieval.* In: Proceedings of the International Conference on Scientific Information (1958). National Academy of Sciences - National Research Council, Washington, D.C., 2, p. 917-935.
11. Masterman, Margaret (1961). *Translation.* Aristotelian Society Supplementary Volume 35, p. 169-216. In: Masterman (2005).
12. Masterman, Margaret (2005). *Language, Cohesion and Form.* Edited by Yorick Wilks. Cambridge University Press, 2005.
13. Parker-Rhodes, A. F. (1956). *Mechanical Translation Program Utilizing an Interlingual Thesaurus.* [Abstract]. In: Report on research: Cambridge Language Research Unit. Mechanical Translation 3, 2, p. 36.
14. Priss, Uta; Old, L. John (2004). *Modelling Lexical Databases with Formal Concept Analysis.* Journal of Universal Computer Science, 10, 8, p. 967-984.

15. Priss, Uta; Old, L. John (2005). *Conceptual Exploration of Semantic Mirrors.* In: Ganter; Godin (eds.), Formal Concept Analysis: Third International Conference, ICFCA 2005, Springer Verlag, LNCS 3403, p. 21-32.
16. Priss, Uta; Old, L. John (2006). *An application of relation algebra to lexical databases.* In: Schaerfe, Hitzler, Ohrstrom (eds.), Conceptual Structures: Inspiration and Application, Proceedings of the 14th International Conference on Conceptual Structures, ICCS'06, Springer Verlag, LNAI 4068, p. 388-400.
17. Priss, Uta; Old, L. John. (2007). *Bilingual Word Association Networks.* In: Priss, Polovina, Hill (eds.), Proceedings of the 15th International Conference on Conceptual Structures, ICCS'07, Springer Verlag, LNAI 4604, p. 310-320.
18. Priss, Uta; Old, L. John. (2008). *Data Weeding Techniques Applied to Roget's Thesaurus.* Knowledge Processing in Practice. (in preparation).
19. Roget, P. M. 1962. *Roget's International Thesaurus.* 3rd Edition Thomas Crowell, New York.
20. Sedelow, S.; Sedelow, W. (1993). *The Concept concept.* Proceedings of the Fifth International Conference on Computing and Information, Sudbury, Ontario, Canada, p. 339-343.
21. Talburt, John R.; Mooney, Donna M. (1989). *The Decomposition of Roget's International Thesaurus into Type-10 Semantically Strong Components.* Proceedings. 1989 ACM South Regional Conference, Tulsa, Oklahoma, p. 78-83.
22. Wilks, Yorick (1998). *Language processing and the thesaurus.* In: Proceedings National Language Research Institute, Tokyo, Japan.
23. Yarowsky, D. (1992). *Word-sense disambiguation using statistical models of Roget's categories trained on large corpora.* In: Proc. COLING92, Nantes, France.