TKK Reports in Information and Computer Science
Espoo 2010 TKK-ICS-R35

CONTEXTUALLY SELF-ORGANIZED MAPS OF CHINESE WORDS, PART II

Teuvo Kohonen

Aalto University
School of Science
and Technology

TKK Reports in Information and Computer Science
Espoo 2010 TKK-ICS-R35

CONTEXTUALLY SELF-ORGANIZED MAPS OF CHINESE WORDS, PART II

Teuvo Kohonen

Aalto University School of Science and Technology
Faculty of Information and Natural Sciences
Department of Information and Computer Science

Aalto-yliopiston teknillinen korkeakoulu
Informaatio- ja luonnontieteiden tiedekunta
Tietojenkäsittelytieteen laitos

Distribution:
Aalto University School of Science and Technology
Faculty of Information and Natural Sciences
Department of Information and Computer Science
PO Box 15400
FI-00076 AALTO
FINLAND
URL: http://ics.tkk.fi
Tel. +358 9 470 01
Fax +358 9 470 23369
E-mail: series@ics.tkk.fi

ISBN 978-952-60-3300-6 (Print)
ISBN 978-952-60-3301-3 (Online)
ISSN 1797-5034 (Print)
ISSN 1797-5042 (Online)
URL: http://lib.tkk.fi/Reports/2010/isbn9789526033013.pdf

AALTO ICS
Espoo 2010

ABSTRACT: In this second publication on the contextual SOMs of Chinese words, a new effect is reported. The SOM was trained by the complete MCRC corpus used in the previous publication. When its hit diagrams were formed using subsets of words of a certain word class with different word frequencies, the hit distribution was found to be a function of this frequency. An explanation of this effect might be that the usage of the words changes with time, and frequent use accelerates this transformation. Therefore, in planning new experiments on the contextually self-organizing word maps, one should be aware of this effect and take it into account in the selection of words to represent the word classes.

KEYWORDS: Chinese text analysis, cognitive linguistics, contextual map, self-organizing map, semantic representation, SOM, word frequency

CONTENTS

Contextually Self-Organized Maps of Chinese Words. Part II: Dependence on Word Frequencies

Teuvo Kohonen

Abstract

In this second work on the contextual SOMs of Chinese words, a new effect has been found and is reported here. The SOM was trained by the complete MCRC corpus used in the previous publication. When its hit diagrams were formed using subsets of words restricted to certain intervals of word frequencies, the hit distribution was found to be a function of the frequency. An explanation of this effect might be that the usage of the words is changed with time, and frequent use accelerates this transformation, which will alter the contexts in which the words appear. Therefore, in planning new experiments on the contextually self-organizing word maps, one should be aware of this effect and take it into account in the selection of words to represent the word classes.

1. The method

The *contextual SOMs* are being used to represent statistical relationships between local contexts (groups of contiguous words) in text corpora, believed to reflect semantic values of the words. It has been found earlier that words of different word classes will be mapped into separate areas on the SOM on the basis of their contexts. In a previous publication (Kohonen, 2010) it was also found that the role of the words as sentence constituents modifies this organization. In the present work it is further found that the mapping depends on the frequencies of the words selected to represent the word classes.

In this work, the construction of contextually self-organizing maps followed the same procedure as described in (Kohonen, 2010); cf. also Refs. 2 and 3. The context around a word, indexed by its position i in the corpus, was defined to consist of five successive words, represented by the coding vectors $\mathbf{r}_{i-2}$, $\mathbf{r}_{i-1}$, $\mathbf{r}_i$, $\mathbf{r}_{i+1}$, and $\mathbf{r}_{i+2}$, respectively.

In order to minimize the effect of the word forms on the contexts and to concentrate on the pure word patterns, the words were encoded and represented by high-dimensional Euclidean vectors with random elements.

For a statistical analysis, each unique word w of the corpus, according to which its context is labeled, is usually described by the *averaged contextual feature* $\mathbf{x}(w)$ of this word, averaged over the corpus. Such features were here formed of five parts, each one being a random-vector representation of a word of the context in the neighborhood of the word w in the text:

$$\mathbf{x}(w) = \mathrm{avg}_i \left([\mathbf{r}_{i-2} \quad \mathbf{r}_{i-1} \quad \mathbf{r}_i \quad \mathbf{r}_{i+1} \quad \mathbf{r}_{i+2}] \mid w \right), \tag{1}$$

where avg $_i$ (***** | *w*) means the average over all positions *i* in the text, on the condition that the contextual feature relating to position *i* belongs to word *w* (i.e., on the condition that $\mathbf{r}_i$ is the random-vector representation of word *w*).

When *constructing* the SOM, the averages of the context vectors **x** *over each unique word w in the corpus* were used as the training data.

The *calibration* of the SOM was carried out using similarly defined input vectors, but averaging the input vectors *only over the words w of a particular word class* (such as general adjectives). Moreover, when the effect of word frequency on the hit diagrams was studied, in the test *only such words (e.g. general adjectives) were included that occurred in the corpus only a specified number of times*, say, 10 to 100 times.

It has further to be mentioned that only such contexts were accepted to the training and testing sets of the SOM inputs that did not contain any punctuation marks or other non-textual symbols.

For computational reasons, the random elements of the **r** vectors were defined by a special hash-coding process described in (Kohonen, 2010). By this method it also became possible to define the dimensionalities of the **r** vectors individually, as a function of their distances from the middle word in the context**.** The dimensionality of the middle vector $\mathbf{r}_i$ was selected as 50. The dimensionalities of $\mathbf{r}_{i-1}$ and $\mathbf{r}_{i+1}$ were taken equal to 200, and those of $\mathbf{r}_{i-2}$ and $\mathbf{r}_{i+2}$ equal to 100, respectively.

2. Hit diagrams

The text corpus used in this work, called the MCRC (Modern Chinese Research Corpus) (Sun el al., 1996), is an electronic collection of text material from newspapers, novels, magazines, TV shows, folktales, and other text material from modern Chinese media. It contains about 1’500’000 words provided with linguistic classification of the words by Dr. Hongbin Xin.

In the following, each linguistic word class will be signified by a special code (e.g., *ag0*, meaning the general adjectives).

2.1 General adjectives (ag0)

In Fig. 1 we have the first *hit diagram*, a histogram of *all general adjectives* (having the linguistic code *ag0*) that occurred in the MCRC corpus. All those testing inputs picked up from ag0 that were closest to a particular model of the SOM were mapped into the corresponding hexagonal bin of the SOM and their number was indicated by the grey level of the bin.

When only “pure” contexts (which did not contain any punctuation marks or other nonlinguistic symbols) were used, the number of remaining unique adjectives (types) was 2002.

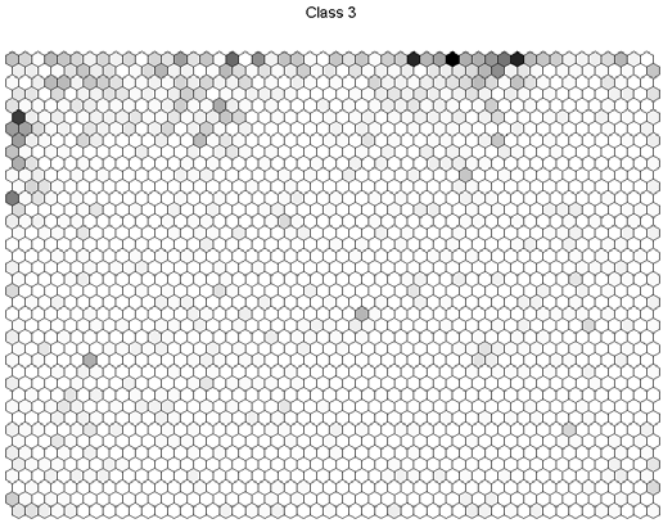


Fig. 1. The hit diagram of 2002 general adjectives (ag0) in the MCRC.

The adjectives in Fig. 1 are mapped onto the top of the SOM into a relatively narrow zone, and further one can discern certain substructures in their distribution, due to the special nature of certain subclasses of adjectives (cf. especially Figs. 3, 4, and 5).

Next we see in Fig.2 the hit diagram constructed *by using only such general adjectives that occurred 1 to 10 times* in the whole MCRC corpus.

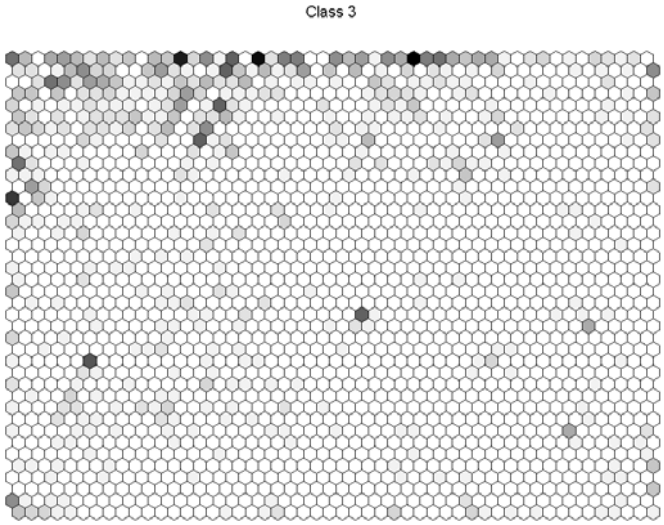


Fig. 2. The hit diagram of those 1059 general adjectives that occurred 1 to 10 times in the MCRC.

The structures of the distributions in Figs.1 and 2 look rather similar, showing that the very rare words, each of which occurs only 1 to 10 times in the corpus, and the number of which is 1059, are the most decisive for the structure of the adjective zone on the SOM.

One particular caution may be necessary, when looking at the graphics of the hit diagrams. In order to be able to compare histograms that contain very different total numbers of hits, one usually normalizes the intensities of the images. The MATLAB graphics does this automatically, unless other options are specified. Then, however, the clusters in low-intensity images are overemphasized and should not be compared directly with clusters in the other images.

Differences with respect to the previous distributions start to show up in Figs. 3 and 4. In Fig. 3, only such test words were used that occurred 10 to 100 times in the corpus. In this diagram, 750 lexical words were included. In Fig. 4, the test words occurred 100 to 1000 times in the corpus, and 217 lexical words were included. The cluster close to the upper right corner is very much emphasized in Figs. 3, 4, and 5. It may be deduced to contain *attributive adjectives* (cf. Sec. 2.11 and an interpretation given in the Discussion).

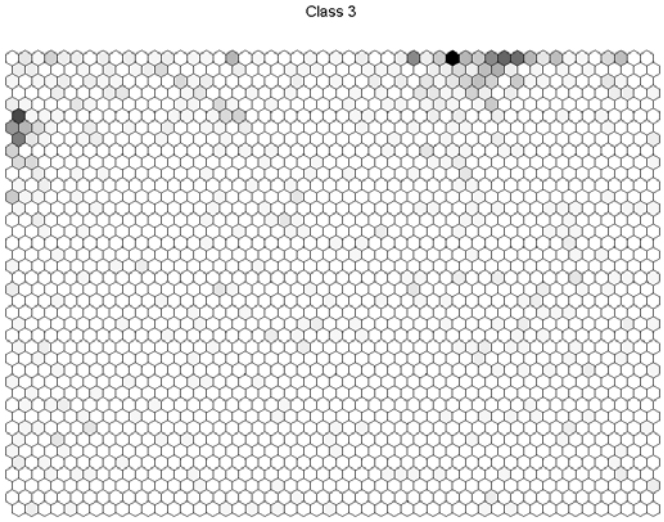


Fig. 3. The hit diagram of those 750 general adjectives that occurred 10 to 100 times in the MCRC.

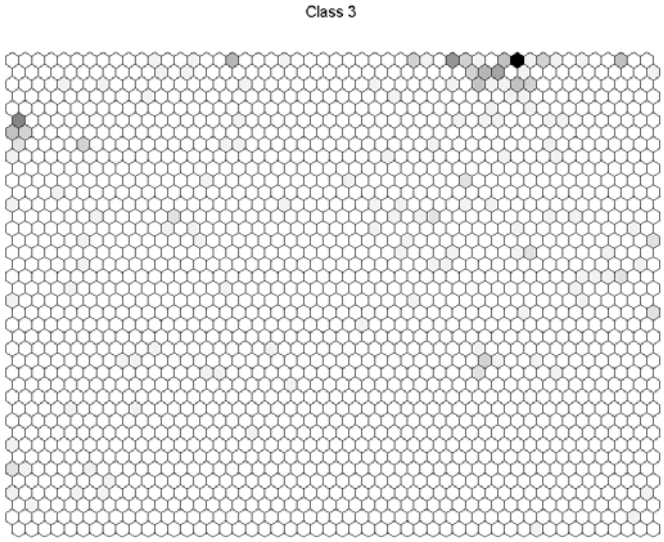


Fig. 4. The hit diagram of those 217 general adjectives that occurred 100 to 1000 times in the MCRC.

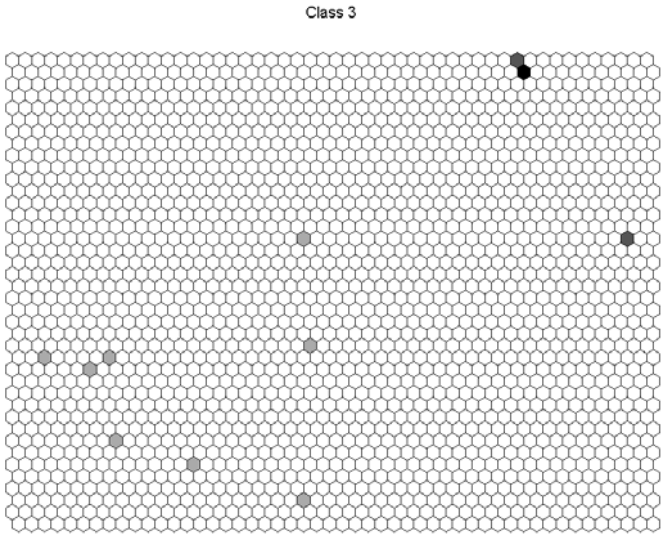


Fig.5. The hit diagram of those 15 general adjectives that occurred 1000 to 10000 times in the MCRC.

In the corpus there were only three lexical words that occurred at least 10000 times. It may be understandable that the contexts in which they occur are no longer statistically comparable with those of the less frequent words, and they may also have a specific semantic role due to their common usage. In Fig. 6 they are scattered more randomly than the words in the previous hit diagrams.

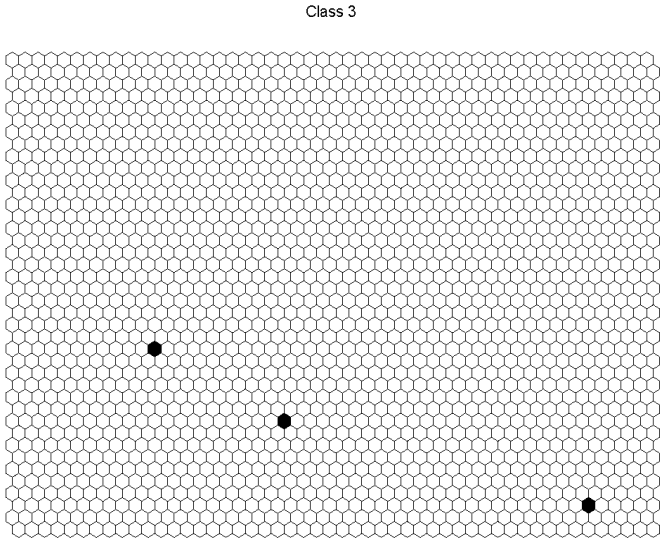


Fig. 6. The hit diagram of those three general adjectives that occurred at least 10000 times in the MCRC.

2.2 Cardinal numbers (mgx)

In all of our earlier experiments on the contextual SOMs, the numerals have been very well clustered. This is obviously due to the fact that the role of the numerals is semantically very simple, well defined, and invariant in various contexts. Now it also seems that the hit diagrams of the numerals are practically independent of the word frequencies, and one may then assume that their role has stayed constant with time. In this subsection we see the hit diagrams of the cardinal numbers.

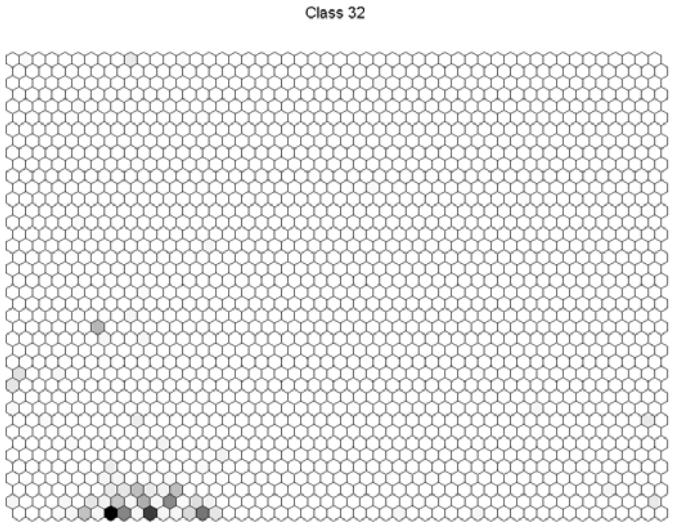


Fig. 7. The hit diagram of 794 cardinal numbers in the MCRC.

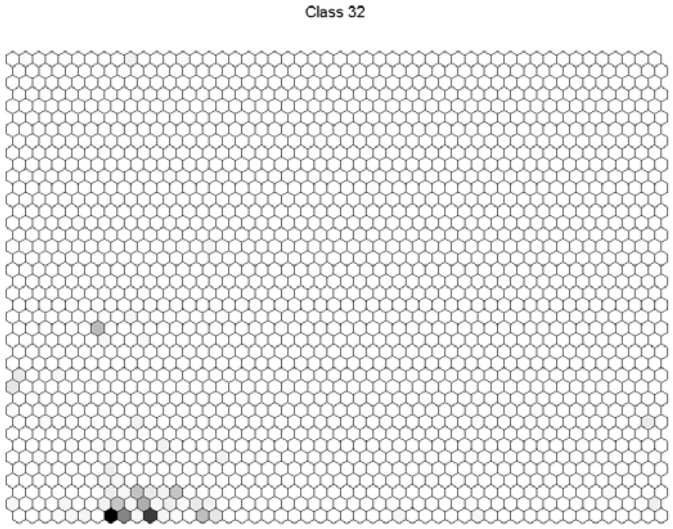


Fig. 8. The hit diagram of those 629 cardinal numbers that occurred 1 to 10 times in the MCRC.

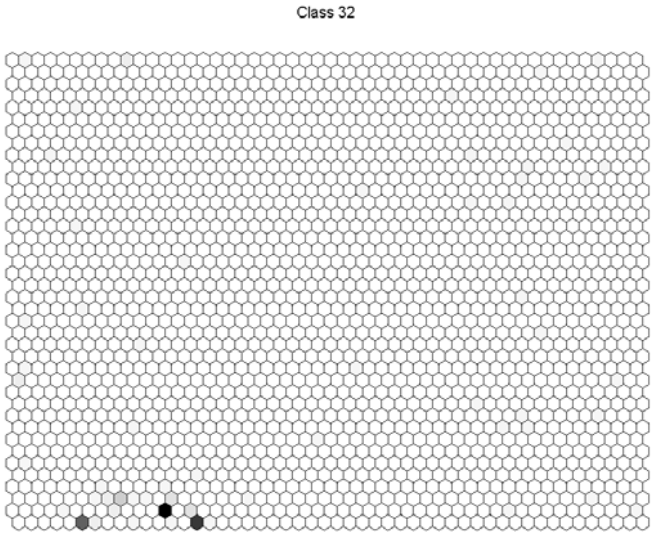


Fig. 9. The hit diagram of those 130 cardinal numbers that occurred 10 to 100 times in the MCRC.

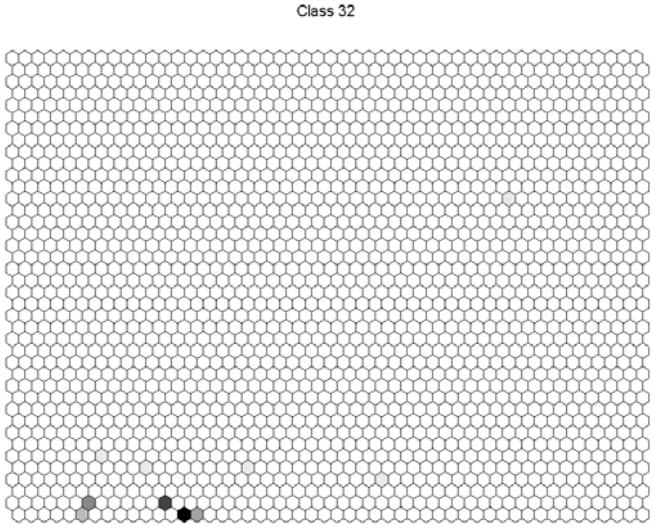


Fig. 10. The hit diagram of those 36 cardinal numbers that occurred 100 to 1000 times in the MCRC.

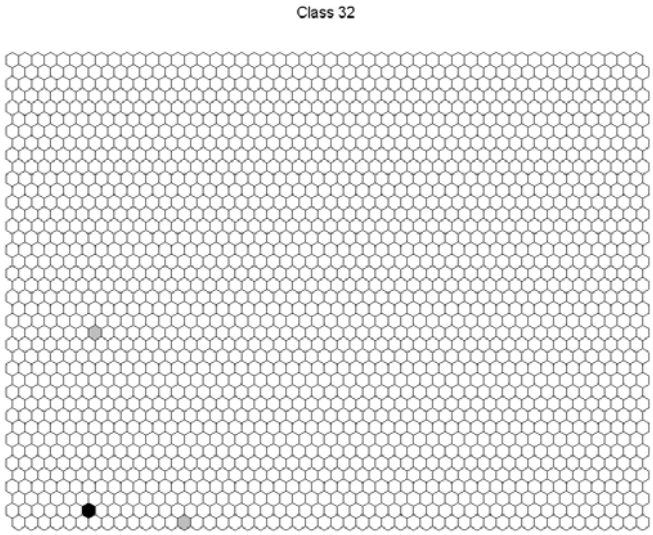


Fig. 11. The hit diagram of those four cardinal numbers that occurred 1000 to 10000 times in the MCRC.

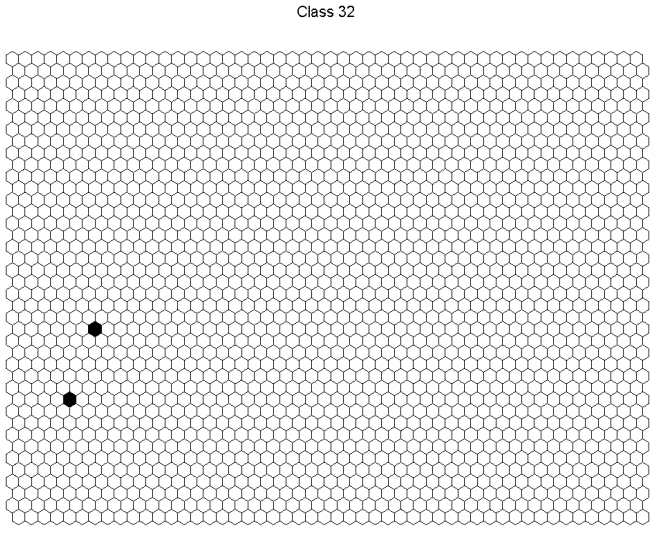


Fig. 12. The hit diagram of those two cardinal numbers that occurred at least 10000 times in the MCRC.

2.3 General nouns (ng0)

In Fig. 13 we have the hit diagram of *all general nouns (*having the linguistic code *ng0)* in the MCRC. It is essentially a very broad belt surrounding an empty area in the middle of the SOM. The latter area, as will be seen later, contains mainly verbs. The number of general nouns in the vocabulary used for this diagram was equal to 10754.

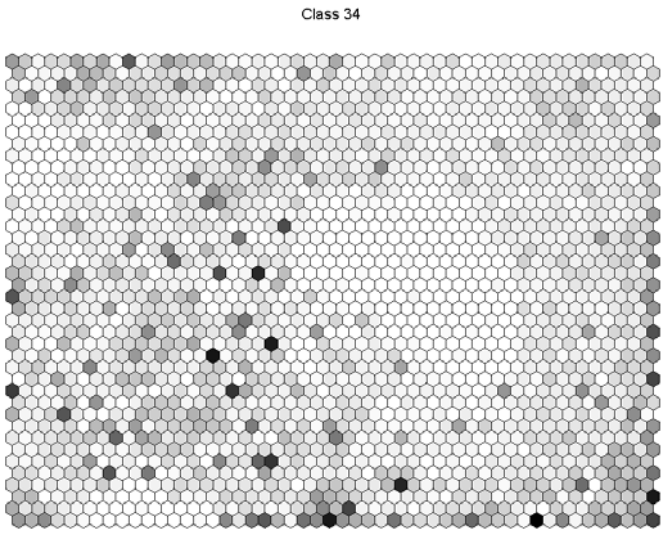


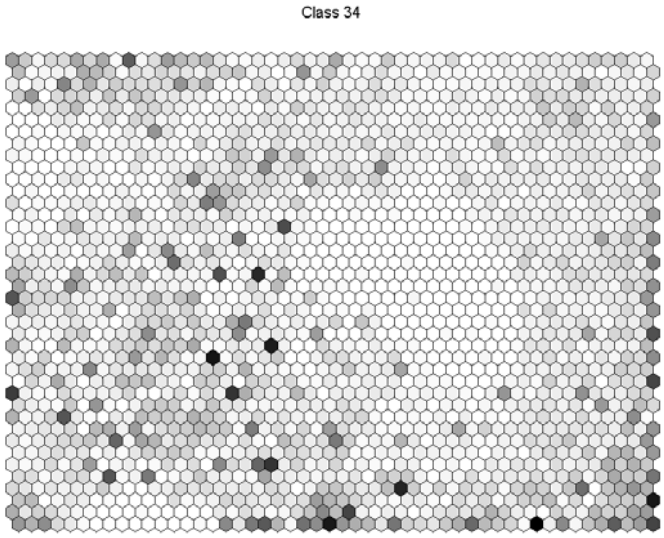
Fig. 13. The hit diagram of 10754 general nouns (ng0) in the MCRC.

At first sight, one might claim that the distribution of the nouns is not clustered at all. The diffuse nature of the image, however, is only virtual. There are several factors that cause the spreading of the hit diagram.

First of all, as will be seen later, the belt is divided into several substructures.

Second, the dependence of the distribution on the frequencies of words selected for testing causes that the distribution is further spread out. With narrowly specified word frequencies, the clustering of this hit diagram is much more constricted, as seen from the images starting with Fig. 16.

In Fig.14 the hit diagram was constructed by using only such general nouns that occurred 1 to 10 times in the whole MCRC corpus. Their number was 6878.



.
Fig. 14. The hit diagram of those 6878 general nouns that occurred 1 to 10 times in the MCRC.

The structures of Figs.13 and 14 look rather similar, apart from some minor differences in their grey levels, showing that the very rare words are decisive also in Fig.13.

Fig. 15 shows the distribution of the 3300 nouns that occurred 10 to 100 times in the corpus. There is a clear shift of the density towards right, compared with Figs. 13 and 14.

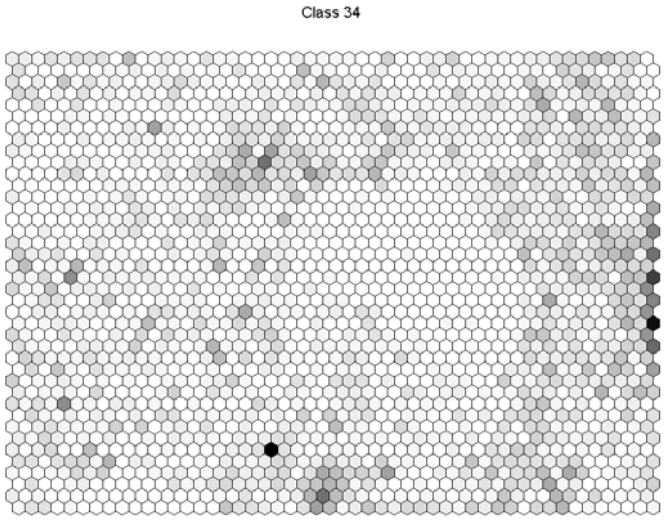


Fig. 15. The hit diagram of those 3300 general nouns that occurred 10 to 100 times in the MCRC.

An even stronger shift of the density towards the upper right corner is visible in Fig.16, which depicts the 733 lexical words that occurred in the corpus 100 to 1000 times. An explanation may be that the semantics of these nouns is now more specific.

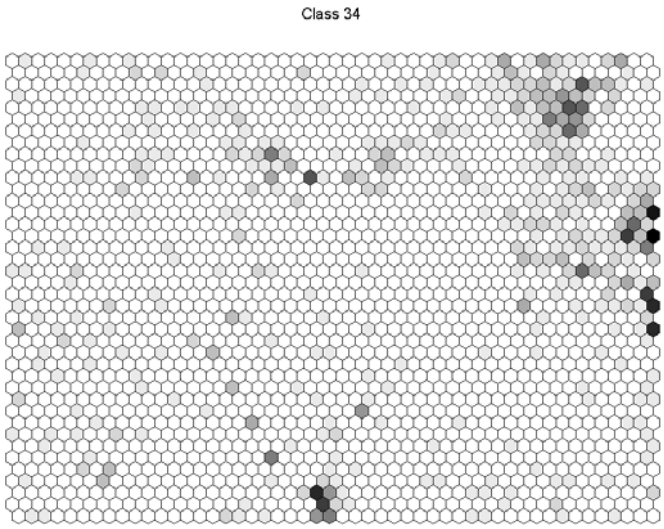


Fig. 16. The hit diagram of those 3300 general nouns that occurred 100 to 1000 times in the MCRC.

The number of lexical words (74) included in the hit diagram of Fig. 17 is so small that one cannot expect to see any statistical forms of the clusters. Nonetheless the hits comply rather well with those of Fig. 16, showing the empty area in the middle.

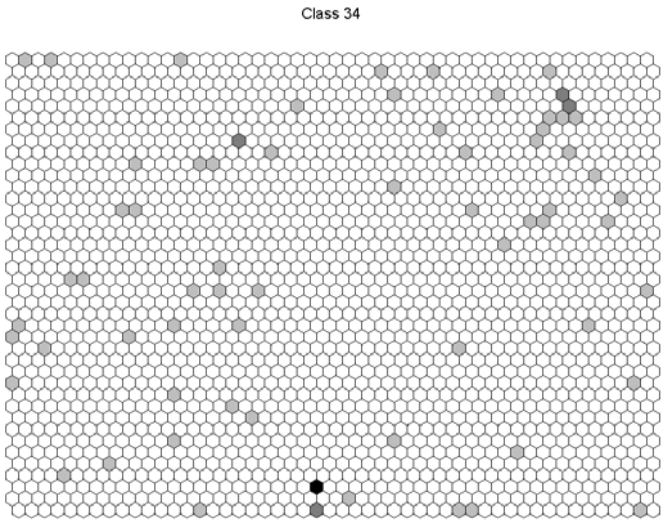


Fig. 17. The hit diagram of those 74 general nouns that occurred 1000 to 10000 times in the MCRC.

The number of lexical words included in the hit diagram of Fig. 18 is only six, and their scattering in the lower left corner cannot be statistically convincing. No conclusions about their contexts can be drawn on the basis of this figure.

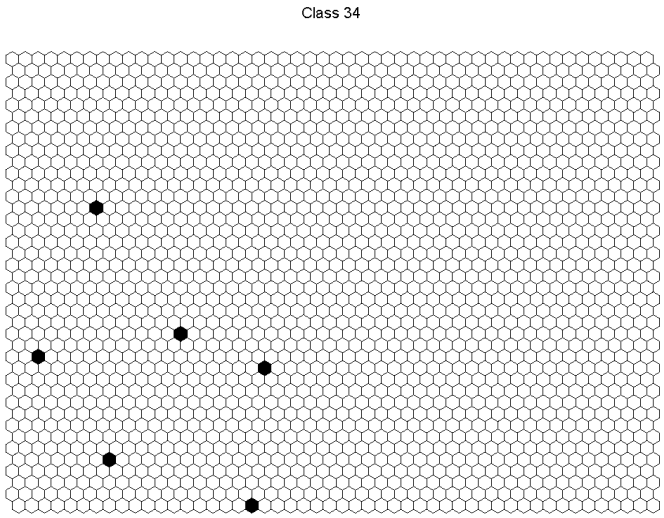


Fig. 18. The hit diagram of those six general nouns that occurred over 10000 times in the MCRC.

2.4 People's names (npm)

The distribution of people's names on the SOM is located roughly in the same area where the rare nouns have their highest density, as seen when comparing Figs. 19 and 14. The cluster in Fig. 19 is narrower and its centroid lies somewhat lower than in Fig. 14.

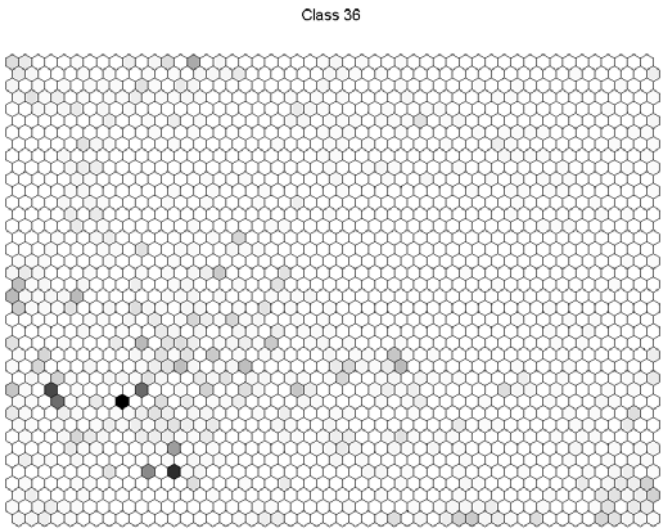


Fig. 19. The hit diagram of 1488 people's names in the MCRC.

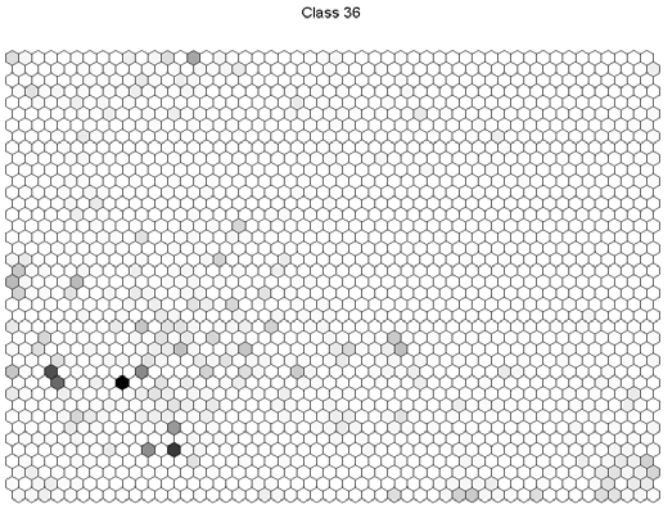


Fig. 20. The hit diagram of those 1209 people's names that occurred 1 to 10 times in the MCRC.

In Fig. 20, only the rarest names (that occurred 1 to 10 times) are shown. The differences between Figs. 19 and 20 are hardly visible, but anyway there exist some faint ones. The main reason for the similarity of Figs. 19 and 20 is that a clear majority (1209/1488) of the names occur only 1 to 10 times in the corpus. There is a high degree of scattering, due to the relatively small number of samples.

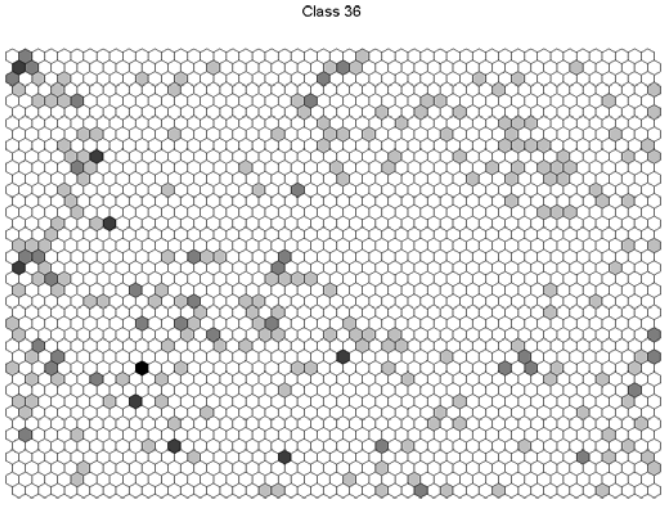


Fig. 21. The hit diagram of those 289 people's names that occurred 10 to 100 times in the MCRC.

In Fig. 21 those 289 names are shown that occurred 10 to 100 times in the corpus. The constellation of Fig. 21 complies with that of the general nouns and also has clearly an

empty area in the middle like, e.g., in Fig. 13. The graphics, however, causes that the low densities seem to be overemphasized.

The next two figures continue the series of people's names. Due to the small number of lexical words included, there exists a high degree of statistical scattering in the hit diagrams, and no conclusions can be drawn from them any longer.

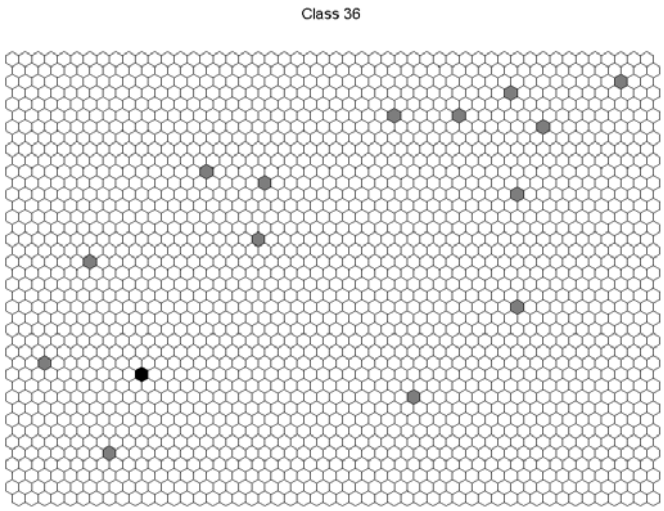


Fig. 22. The hit diagram of those 16 people's names that occurred 100 to 1000 times in the MCRC. Notice that one of the three dots in the lower left corner has the double intensity with respect to the rest.

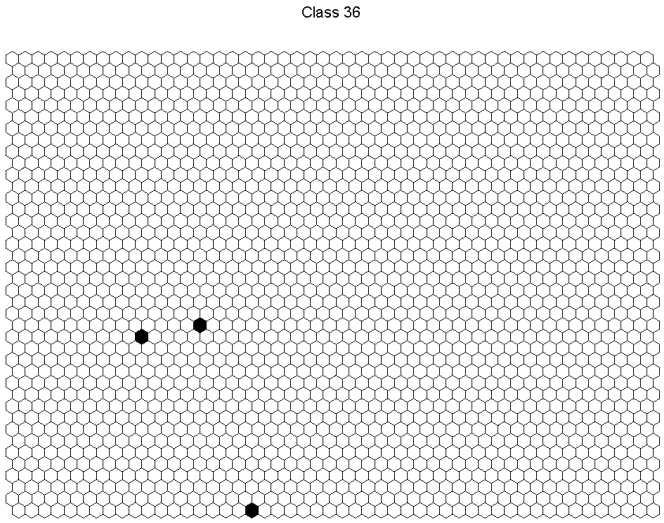


Fig. 23. The hit diagram of those three people's names that occurred 1000 to 10000 times in the MCRC

2.5 Names of places (nps)

The following five hit diagrams of names of places resemble those of people's names, but they are slightly offset from the latter. In Fig. 24 those 1213 names of places are shown that occurred in the corpus.

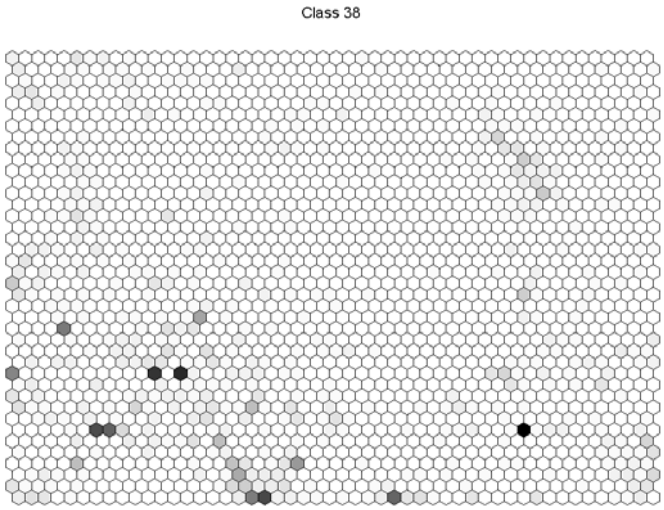


Fig. 24. The hit diagram of 1213 names of places in the MCRC.

In Fig.25 we see the hit diagram constructed by using only such general nouns that occurred 1 to 10 times in the MCRC corpus. Again there is a high degree of similarity between Figs. 24 and 25, whereas the next distributions already will change gradually.

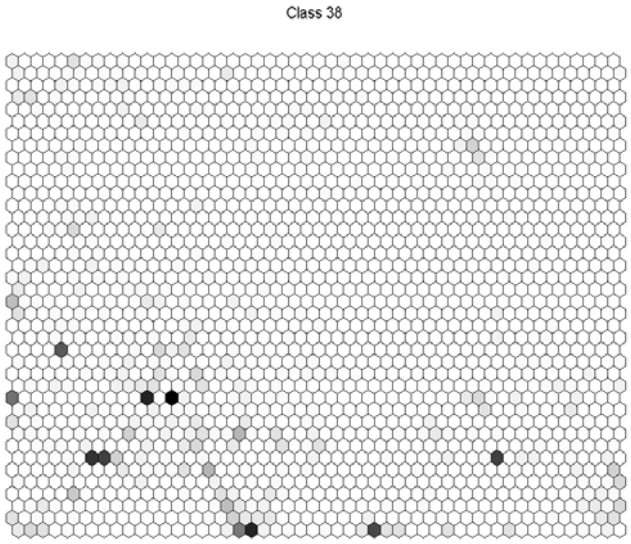


Fig. 25. The hit diagram of those 906 names of places that occurred 1 to 10 times in the MCRC.

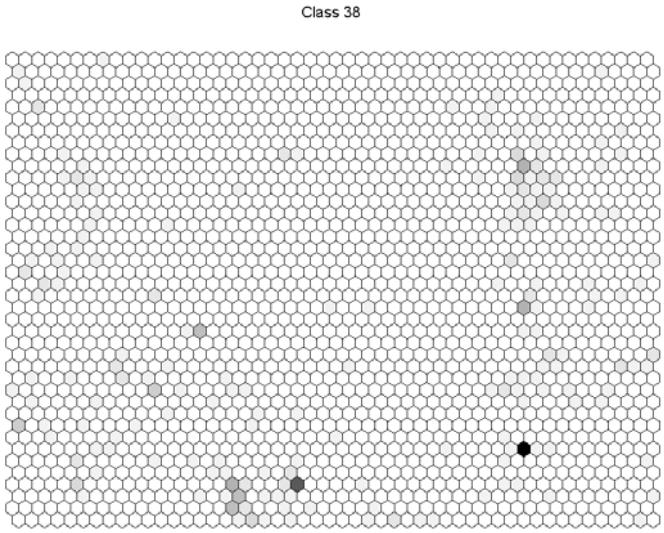


Fig. 26. The hit diagram of those 289 names of places that occurred 10 to 100 times in the MCRC.

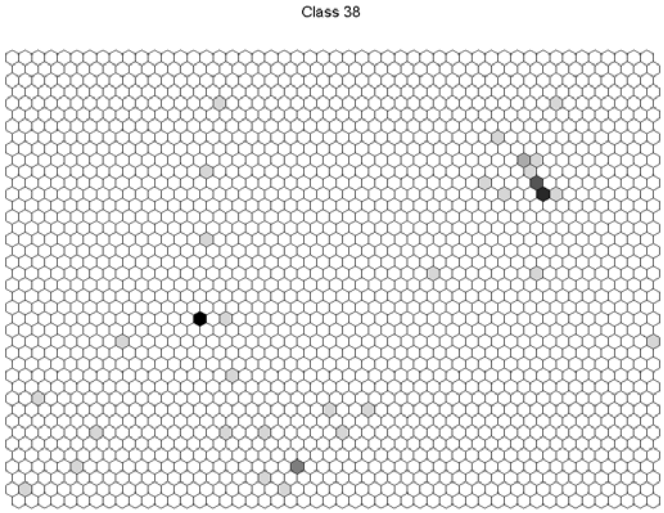


Fig. 27. The hit diagram of those 47 names of places that occurred 100 to 1000 times in the MCRC.

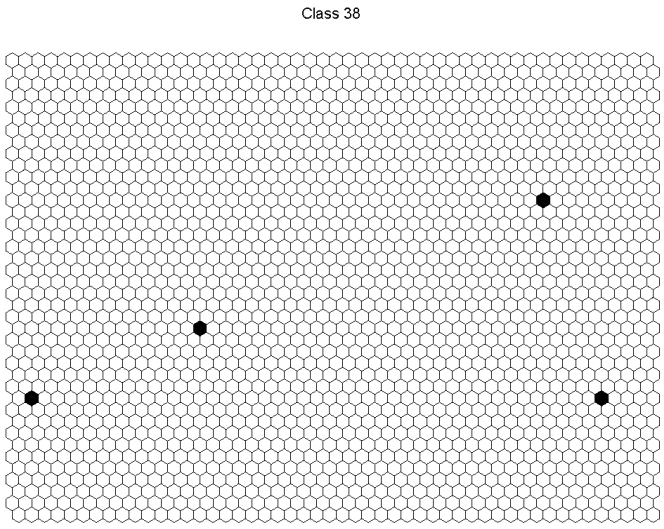


Fig. 28. The hit diagram of those four names of places that occurred 1000 to 10000 times in the MCRC.

2.6 Nouns of time (nt)

The hit diagrams of the nouns that are related to time differ slightly from those of the previous two classes, as seen from their concentration on a vertical zone slightly to the right of the area where the names of persons are concentrated.

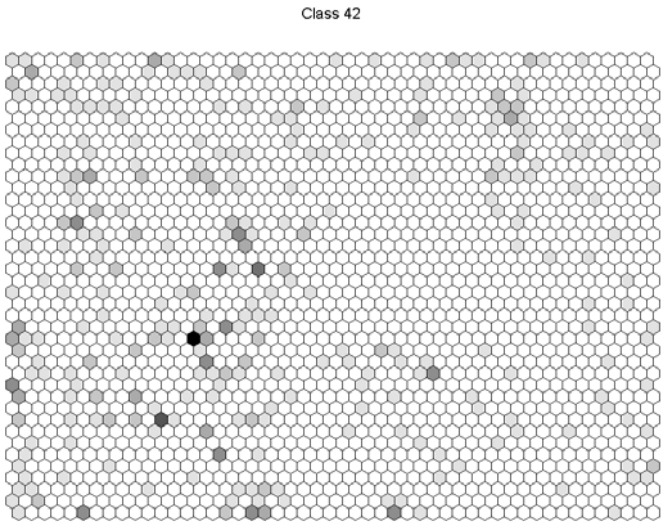


Fig. 29. The hit diagram of 410 nouns of time in the MCRC.

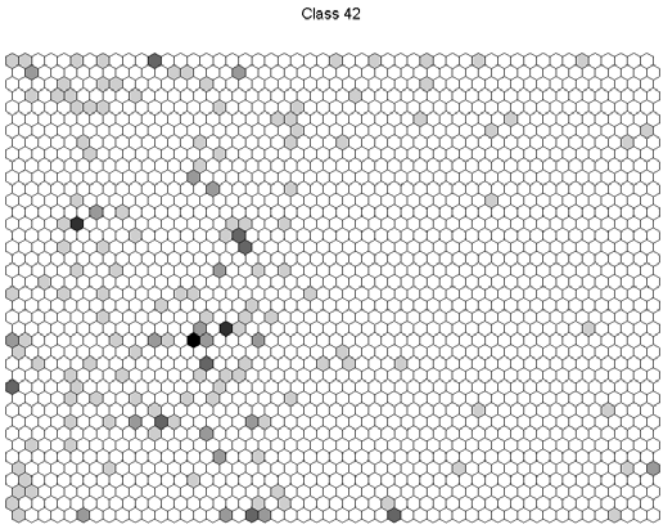


Fig. 30. The hit diagram of those 198 nouns of time that occurred 1 to 10 times in the MCRC.

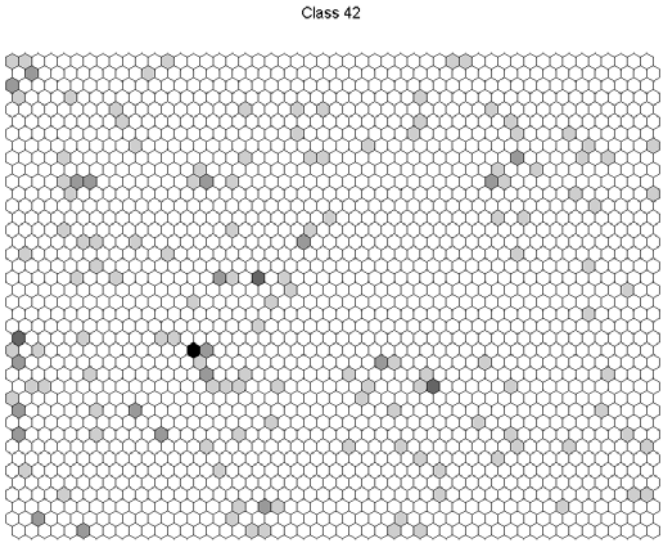


Fig. 31 The hit diagram of those 177 nouns of time that occurred 10 to 100 times in the MCRC.

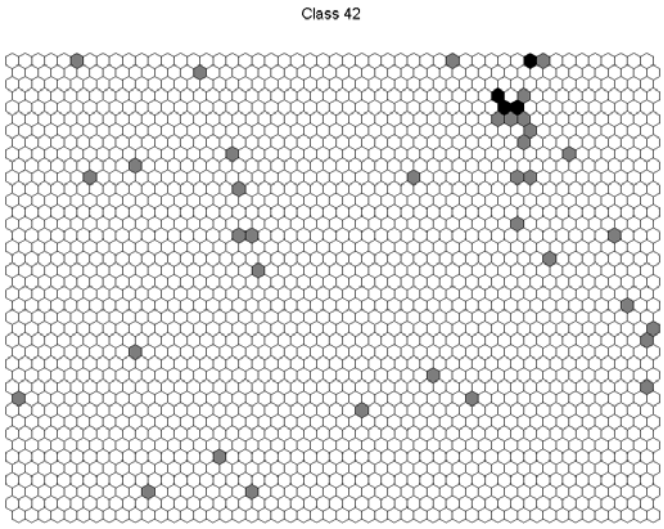


Fig. 32 The hit diagram of those 44 nouns of time that occurred 100 to 1000 times in the MCRC.

The cluster in the upper right corner of Fig. 32 is strongly emphasized, apparently, however, because the automatic normalization of the grey levels causes a distortion when there are only few hits per model (cf. Discussion). No other explanation is easily suggestible.

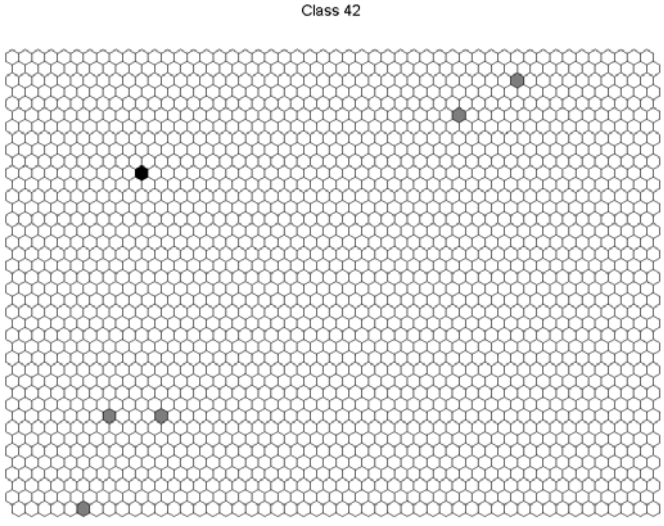


Fig. 33. The hit diagram of those seven nouns of time that occurred 1000 to 10000 times in the MCRC.

2.7 Verbs without objects (vw0)

This class, related to the *intransitive verbs* in some other languages, has the main cluster in the middle of the SOM where the nouns have a round empty area. In some diagrams there is a singular cluster at the top of the map where the corresponding verbs probably have a special nature.

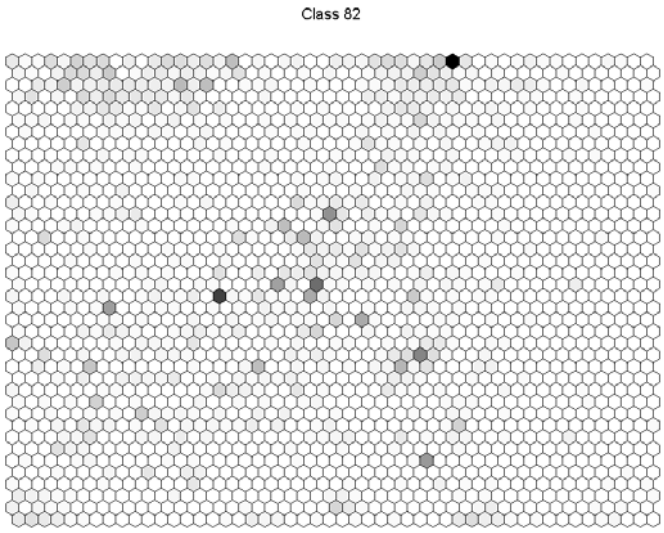


Fig. 34. The hit diagram of 5053 verbs without objects in the MCRC.

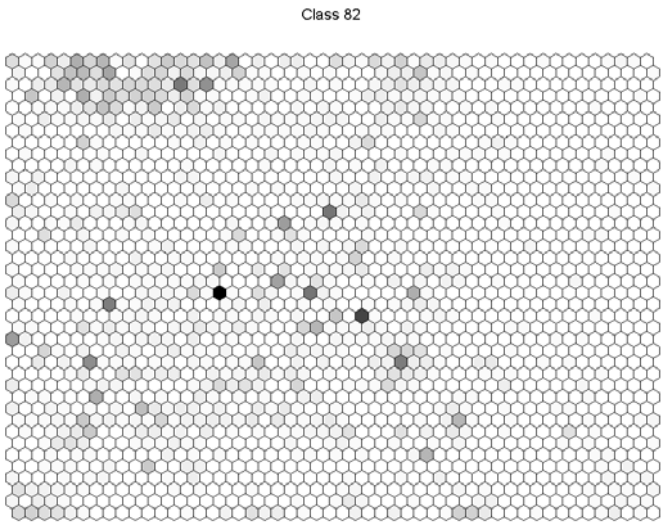


Fig. 35. The hit diagram of those 2518 verbs without objects that occurred 1 to 10 times in the MCRC.

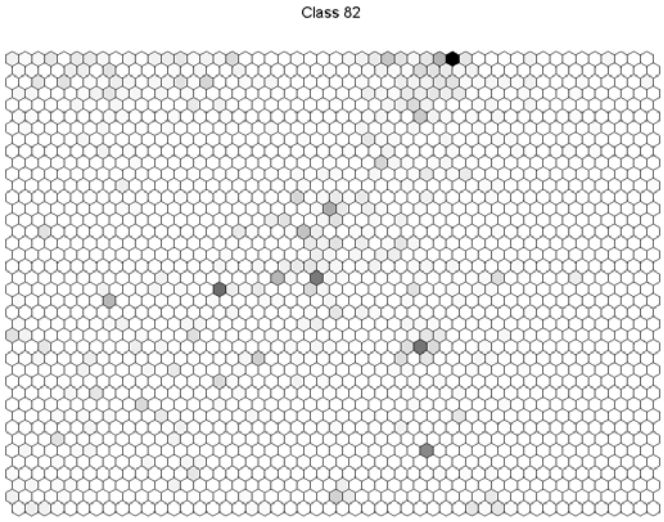


Fig. 36. The hit diagram of those 2057 verbs without objects that occurred 10 to 100 times in the MCRC.

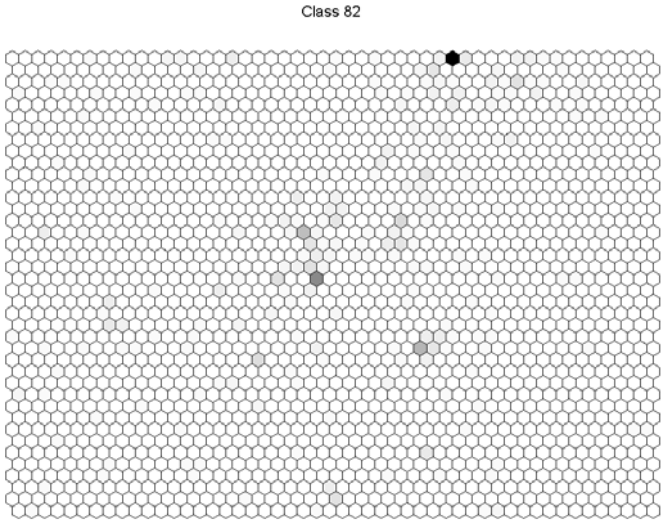


Fig. 37. The hit diagram of those 525 verbs without objects that occurred 100 to 1000 times in the MCRC.

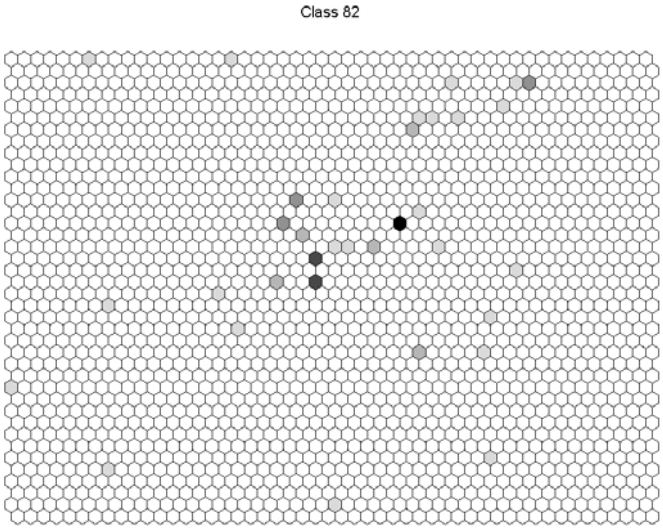


Fig. 38. The hit diagram of those 89 verbs without objects that occurred 1000 to 10000 times in the MCRC.

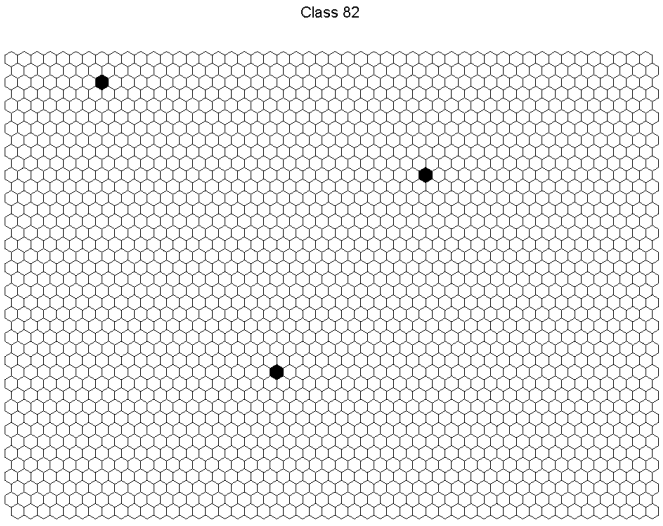


Fig. 39. The hit diagram of those five verbs without objects that occurred at least 10000 times in the MCRC.

2.8 Verbs followed by nouns (vwn)

The verbs of this class are related to the *transitive verbs* in some other languages. Their hit diagrams are very well clustered in the middle of the SOM where the nouns have an empty area, and they are almost invariant with respect to the frequency of words.

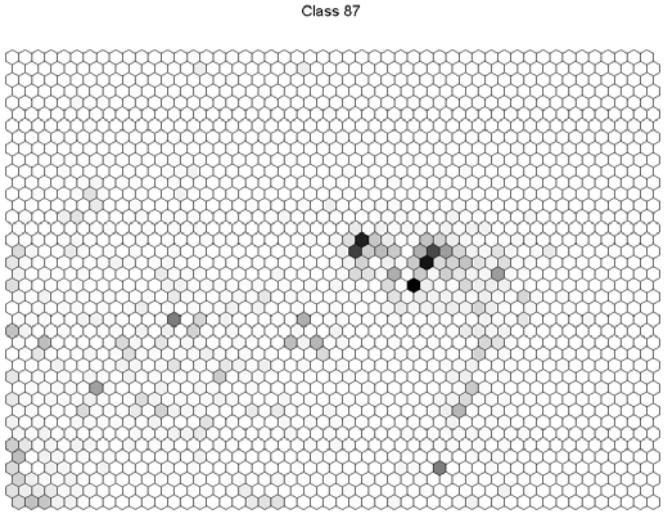


Fig. 40. The hit diagram of 3691 verbs followed by nouns in the MCRC.

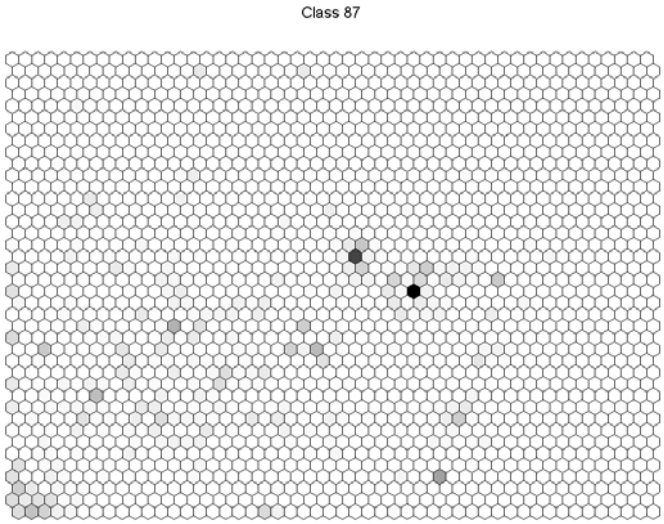


Fig. 41. The hit diagram of those 1577 verbs followed by nouns that occurred 1 to 10 times in the MCRC.

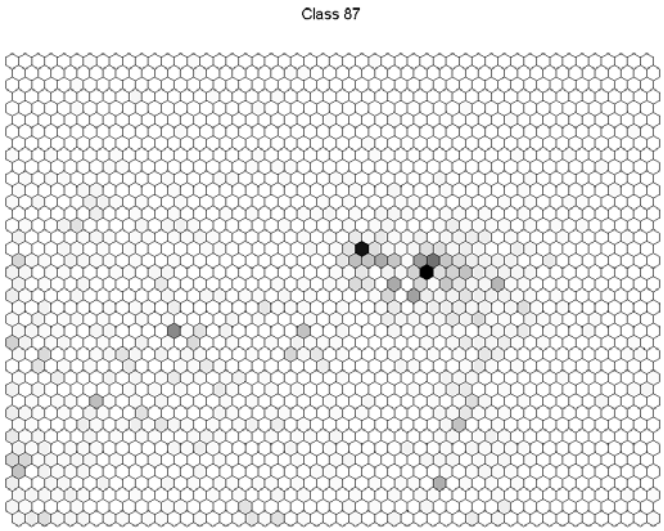


Fig. 42. The hit diagram of those 1642 verbs followed by nouns that occurred 10 to 100 times in the MCRC.

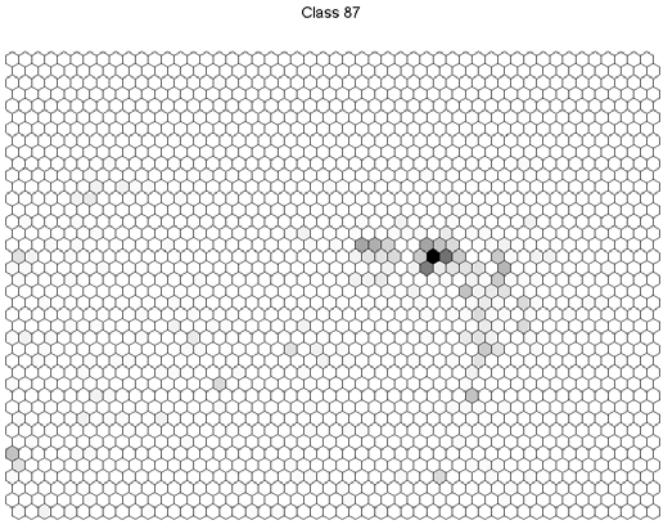


Fig. 43. The hit diagram of those 489 verbs followed by nouns that occurred 100 to 1000 times in the MCRC.

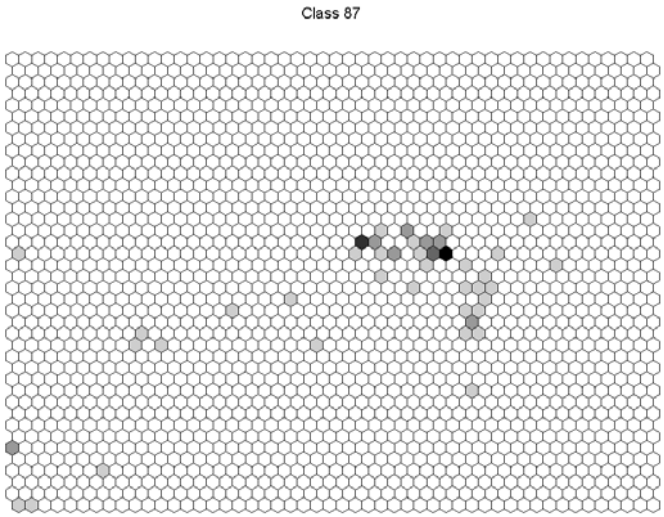


Fig. 44. The hit diagram of those 59 verbs followed by nouns that occurred 1000 to 10000 times in the MCRC.

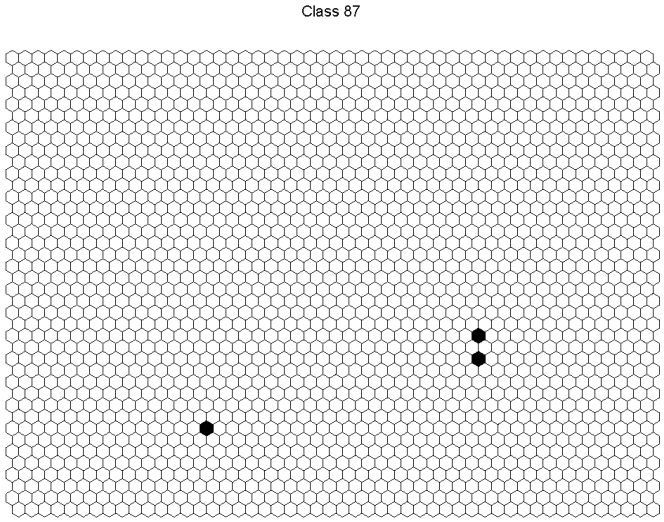


Fig. 45. The hit diagram of those three verbs followed by nouns that occurred at least 10000 times in the MCRC.

It is interesting to note that the hit diagrams of the verbs followed by nouns are so well clustered in the middle of the area where the nouns have an empty space. This may be due to the fact that these verbs are then almost optimally close to the majority of the nouns.

2.9 Verbs as modifying components in noun phrases (vtp)

The hit diagrams of this class are clearly different from any of those of the previous two classes. In particular, *they have an empty area in the middle of the map like the nouns, so that their roles as sentence constituents resemble those of the nouns.*

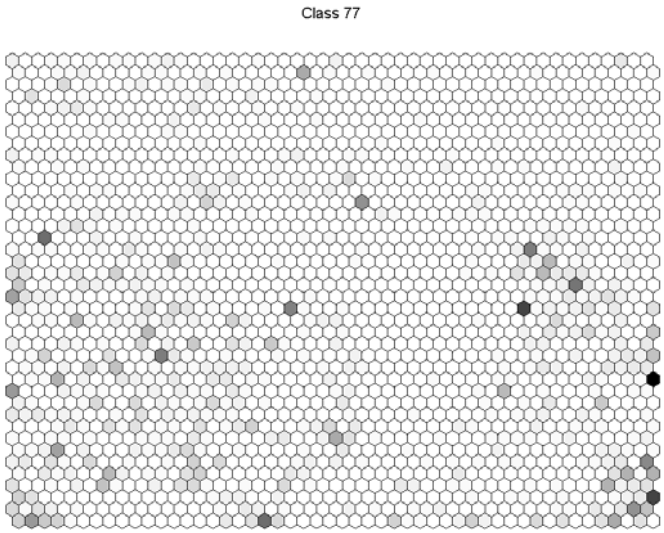


Fig. 46. The hit diagram of 1534 verbs as modifying components in noun phrases in the MCRC.

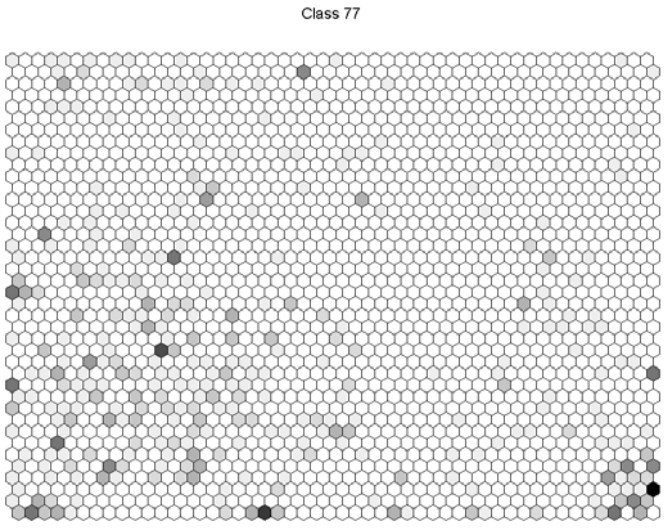


Fig. 47. The hit diagram of those 605 verbs as modifying components in noun phrases that occurred 1 to 10 times in the MCRC.

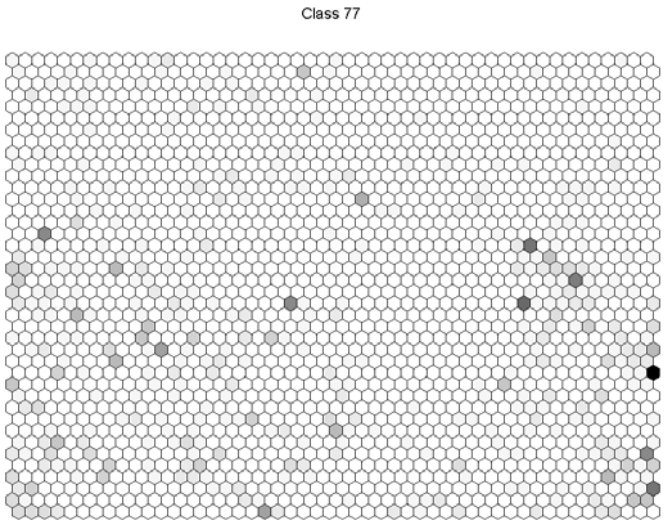


Fig. 48. The hit diagram of those 727 verbs as modifying components in noun phrases that occurred 10 to 100 times in the MCRC.

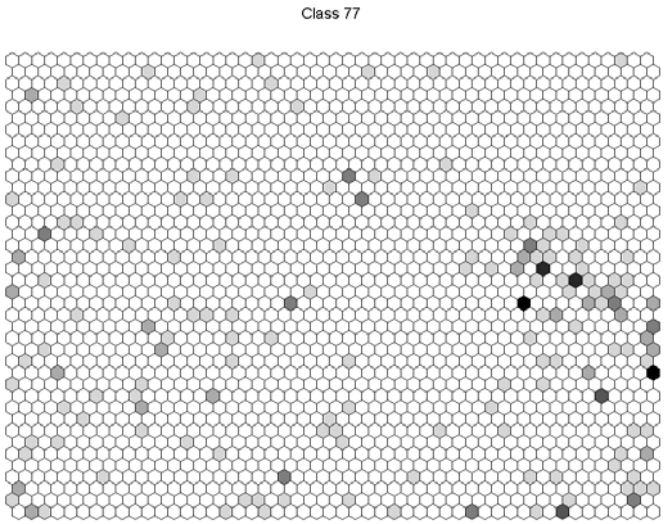


Fig. 49. The hit diagram of those 222 verbs as modifying components in noun phrases that occurred 100 to 1000 times in the MCRC.

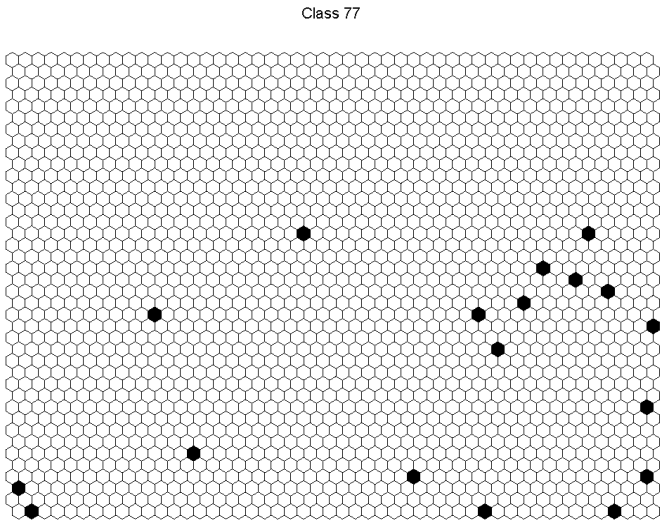


Fig. 50. The hit diagram of those 18 verbs as modifying components in noun phrases that occurred 1000 to 10000 times in the MCRC.

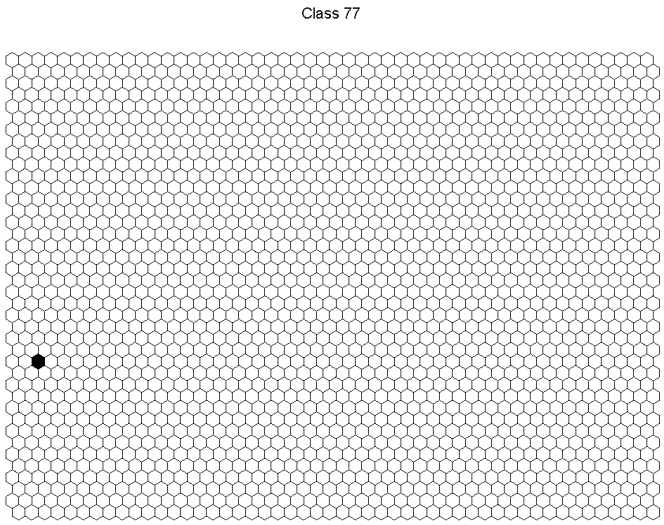


Fig. 51. The hit diagram of one verb as a modifying component in a noun phrase that occurred at least 10000 times in the MCRC.

2.10 Verbs as the core of a noun phrase (vtx)

The hit diagrams of this class, let alone the last figure, are concentrated around the lower right corner. As they are clearly different from any of those of the other verb classes, one has to conclude that the semantic nature of this class is also very different from the rest.

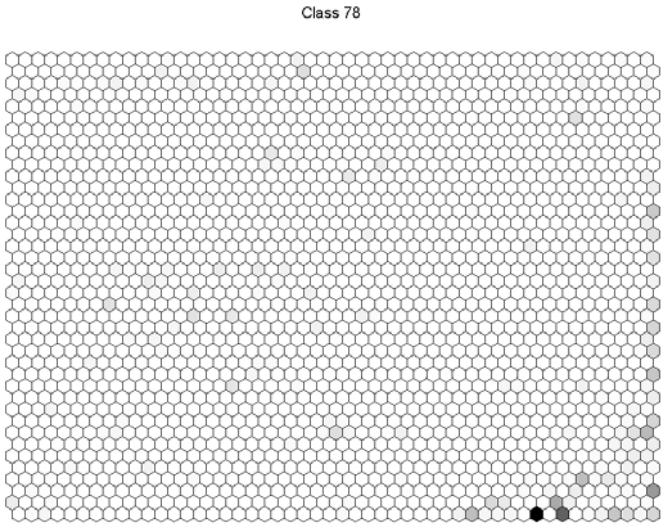


Fig. 52. The hit diagram of 1035 verbs as the core of a noun phrase in the MCRC.

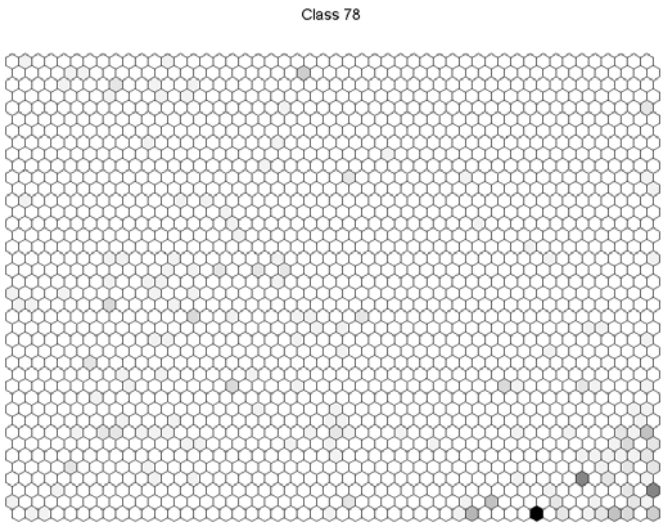


Fig. 53. The hit diagram of those 253 verbs as the core of a noun phrase that occurred 1 to 10 times in the MCRC.

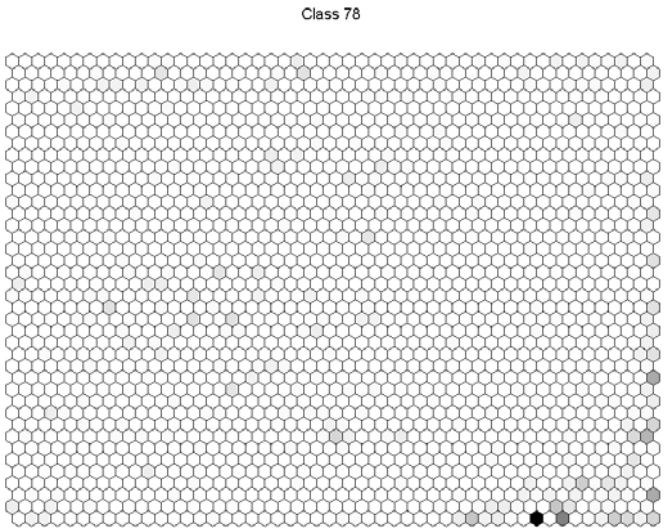


Fig. 54. The hit diagram of those 567 verbs as the core of a noun phrase that occurred 10 to 100 times in the MCRC.

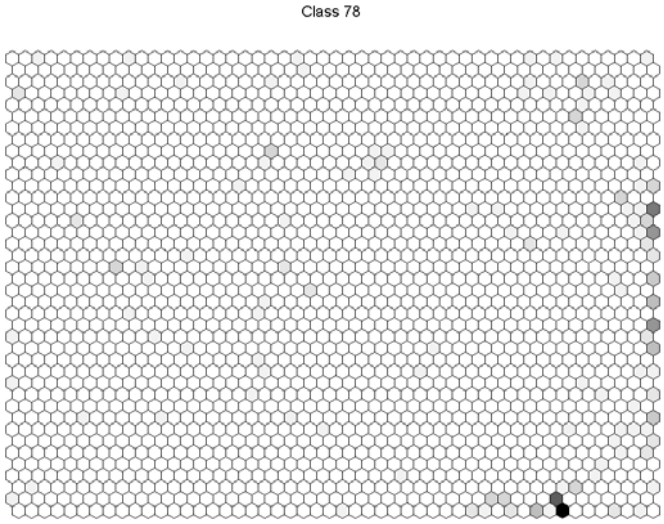


Fig. 55 The hit diagram of those 220 verbs as the core of a noun phrase that occurred 100 to 1000 times in the MCRC.

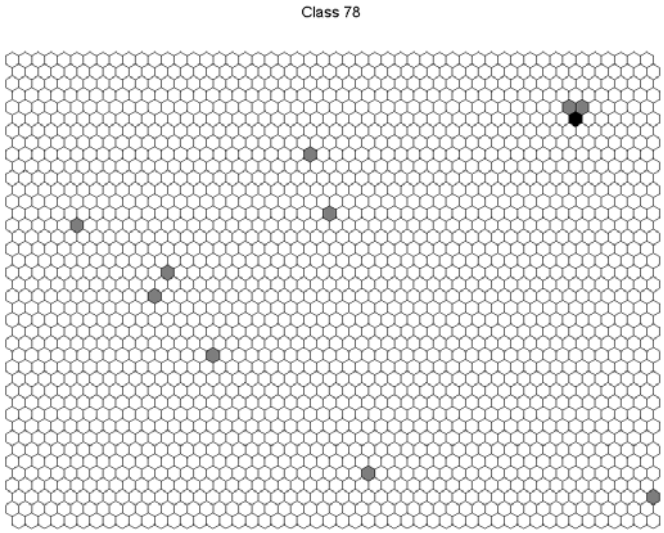


Fig. 56 The hit diagram of those 12 verbs as the core of a noun phrase that occurred 1000 to 10000 times in the MCRC.

2.11 Pronouns as attributive adjectives (rn)

This is a very interesting class, because many of its hit diagrams have their main cluster at the top, to the right of the middle, in the same place where some adjectives also have a strong cluster (cf. Figs. 3, 4, and 5). The latter thus probably represents attributive adjectives.

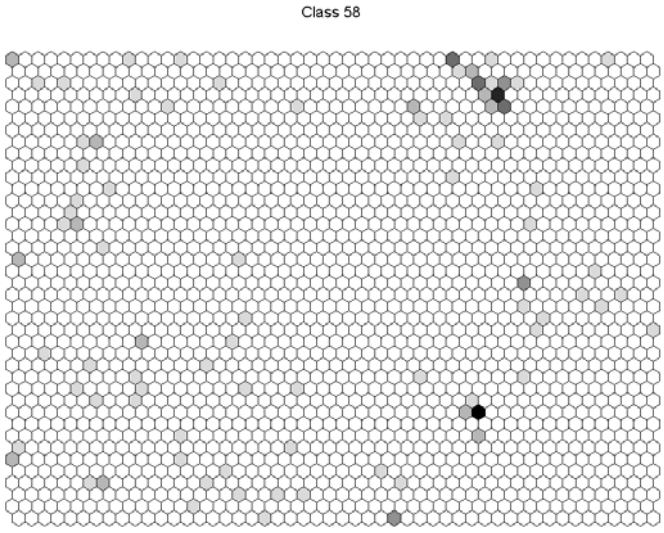


Fig. 57. The hit diagram of 125 pronouns as adjective attributes in the MCRC.

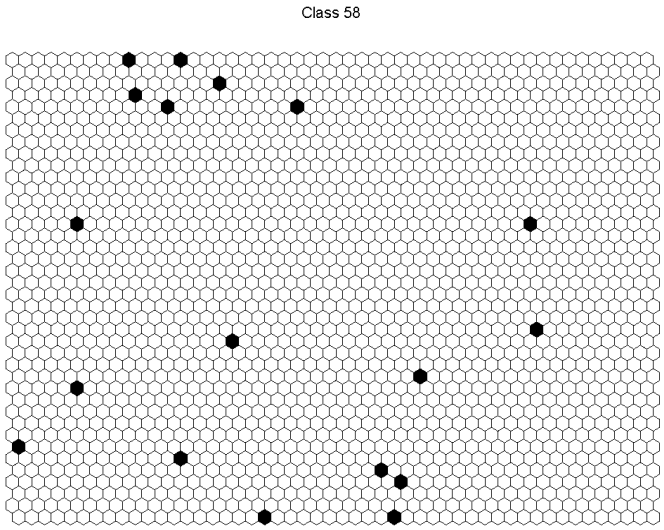


Fig.58. The hit diagram of those 18 pronouns as adjective attributes that occurred 1 to 10 times in the MCRC.

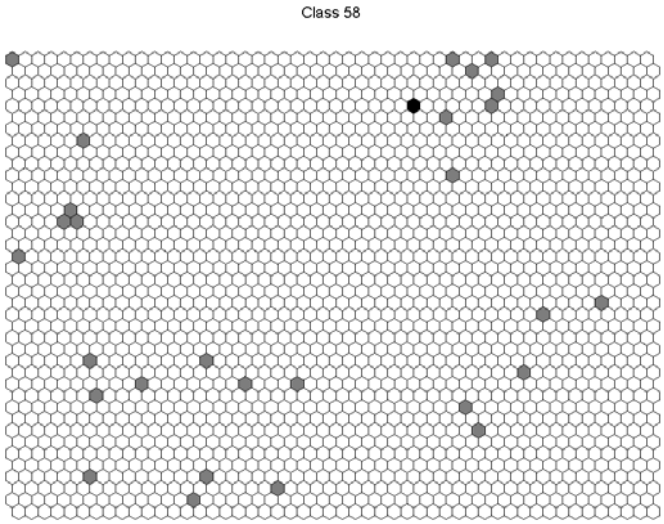


Fig.59. The hit diagram of those 30 pronouns as adjective attributes that occurred 10 to 100 times in the MCRC.

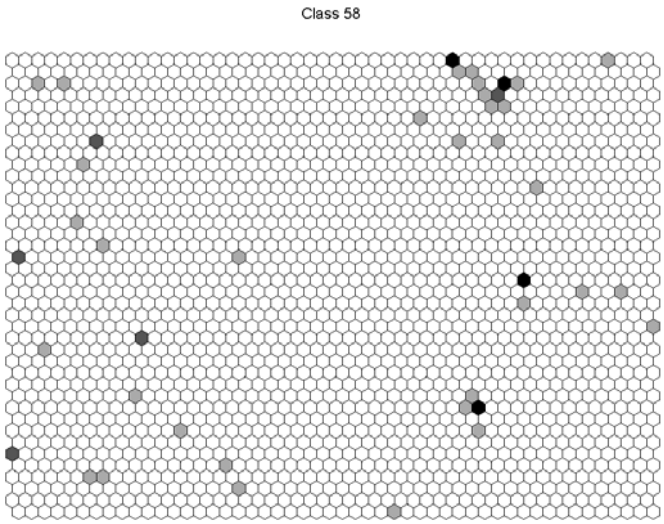


Fig.60. The hit diagram of those 55 pronouns as adjective attributes that occurred 100 to 1000 times in the MCRC.

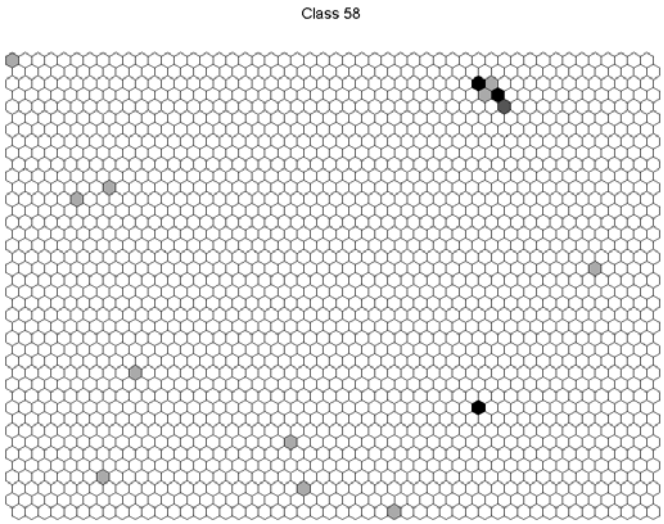


Fig.61. The hit diagram of those 22 pronouns as adjective attributes that occurred 1000 to 10000 times in the MCRC.

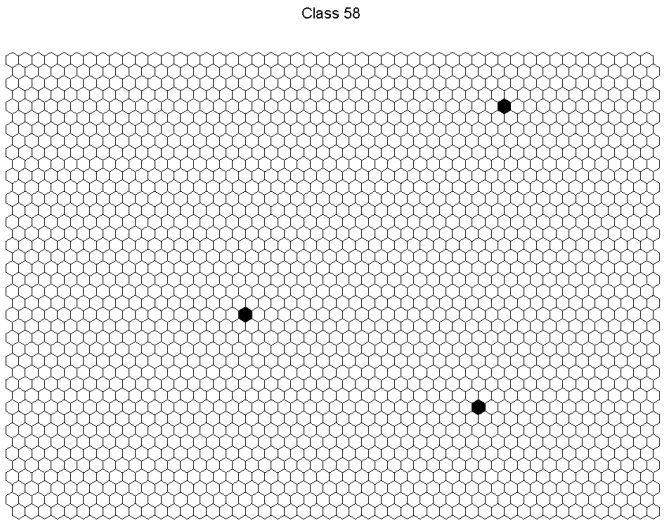


Fig.62. The hit diagram of those three pronouns as adjective attributes that occurred at least 10000 times in the MCRC.

2.12 Pronouns as subjects or objects (rs)

This class has hit diagrams similar to those of the *nouns*, especially *having an empty area in the middle of the SOM, indicating that the roles of these words as sentence constituents are similar to those of the nouns*.

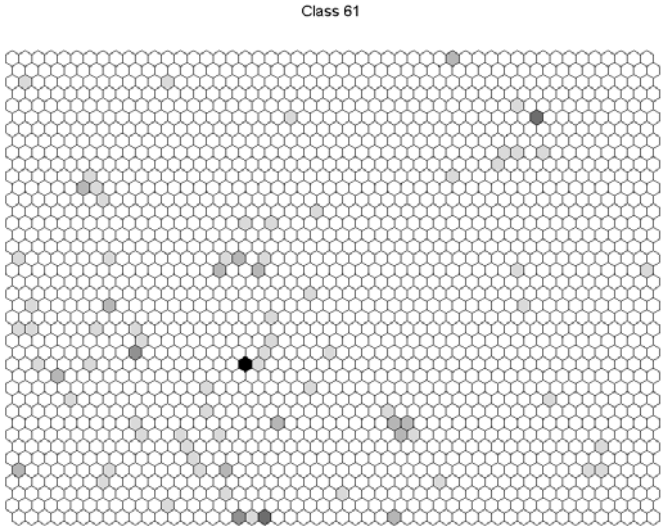
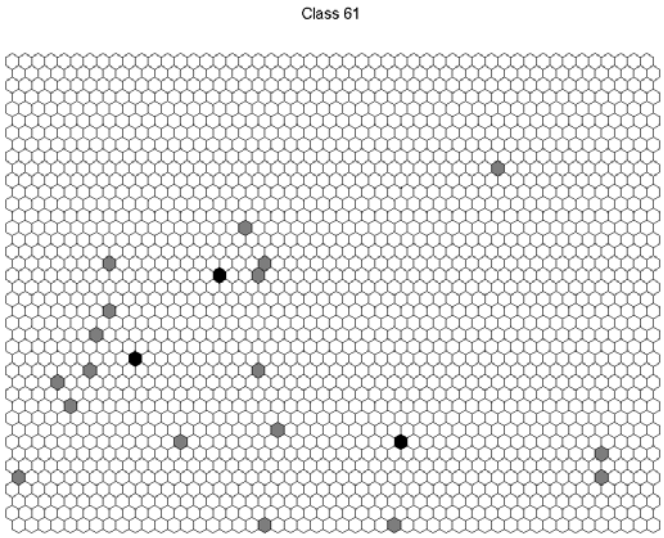
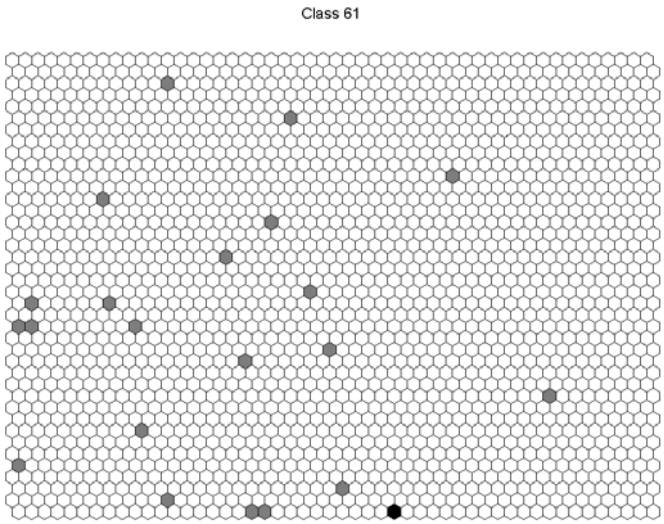


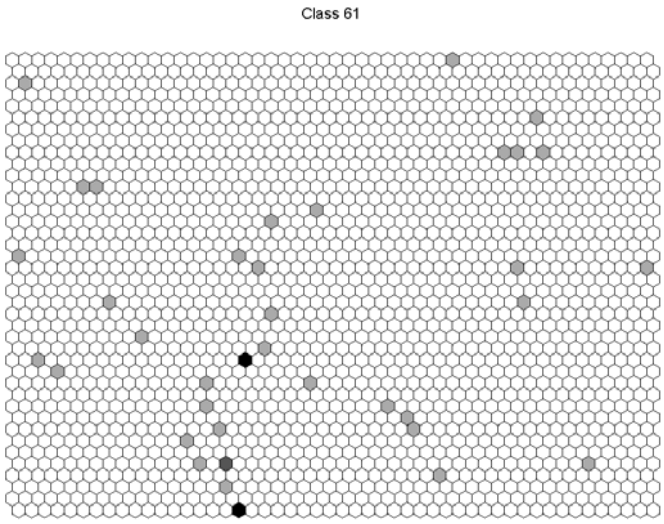
Fig. 63. The hit diagram of those 109 pronouns as subjects or objects that occurred in the MCRC.



Fig, 64. The hit diagram of those 27 pronouns as subjects or objects that occurred 1 to 10 times in the MCRC.



Fig, 65. The hit diagram of those 23 pronouns as subjects or objects that occurred 10 to 100 times in the MCRC.



Fig, 66. The hit diagram of those 42 pronouns as subjects or objects that occurred 100 to 1000 times in the MCRC.

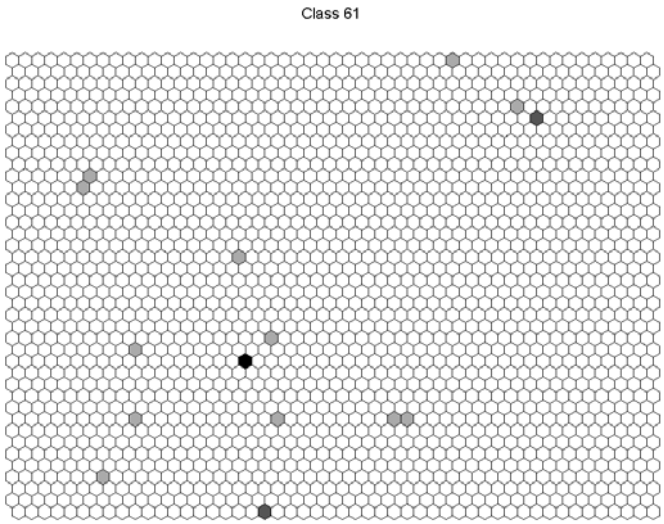


Fig. 67. The hit diagram of those 19 pronouns as subjects or objects that occurred 1000 to 10000 times in the MCRC.

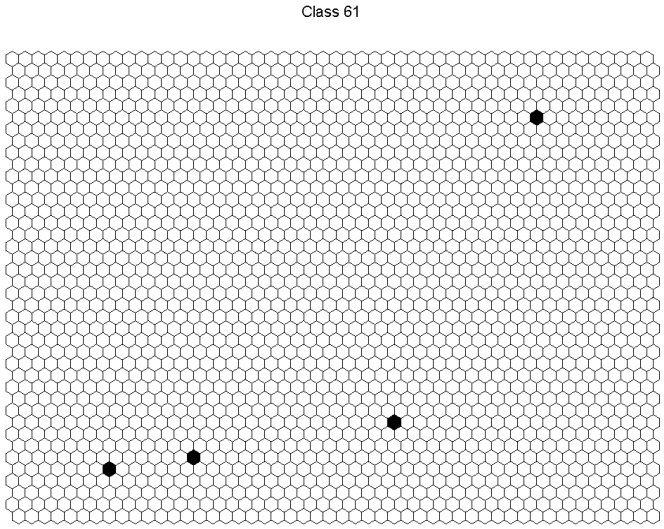


Fig. 68. The hit diagram of those four pronouns as subjects or objects that occurred at least 10000 times in the MCRC.

2.13 General prepositions (pg)

Especially the hit diagrams of all of the general prepositions, shown in Fig. 69, as well as the partial hit diagrams shown in Figs. 72 through 74 seem to be clustered into a narrow area to the right of the middle of the SOM. The prepositions shown in Figs. 70 and 71 have a different distribution. Maybe there is a certain subclass of exceptional cases in Figs. 70 and 71.

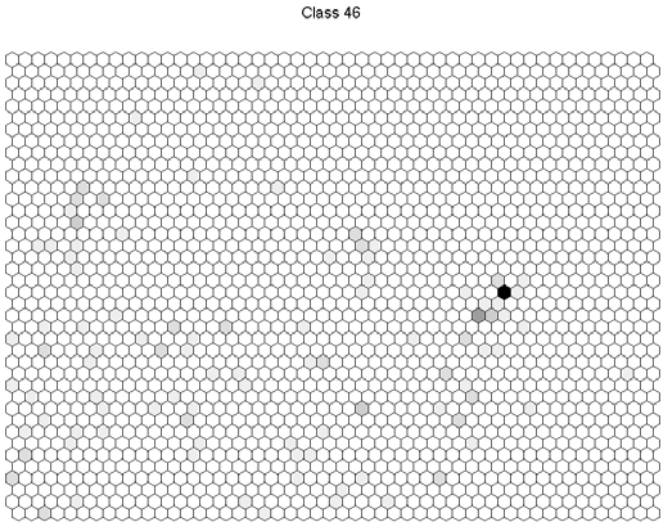


Fig. 69. The hit diagram of those 143 general prepositions that occurred in the MCRC.

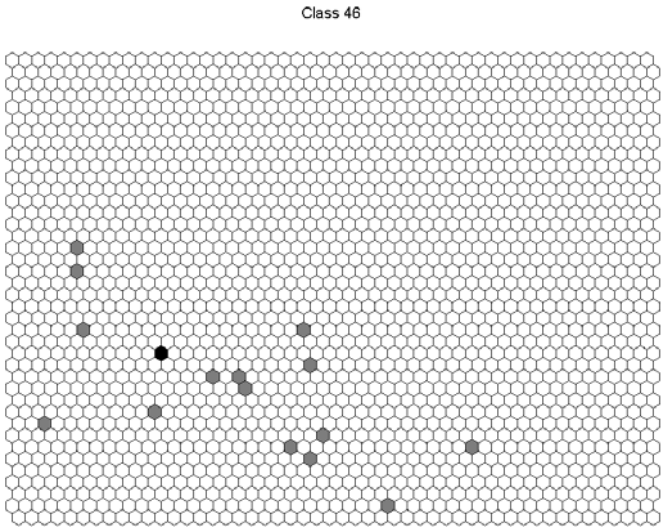


Fig. 70. The hit diagram of those 17 general prepositions that occurred 1 to 10 times in the MCRC.

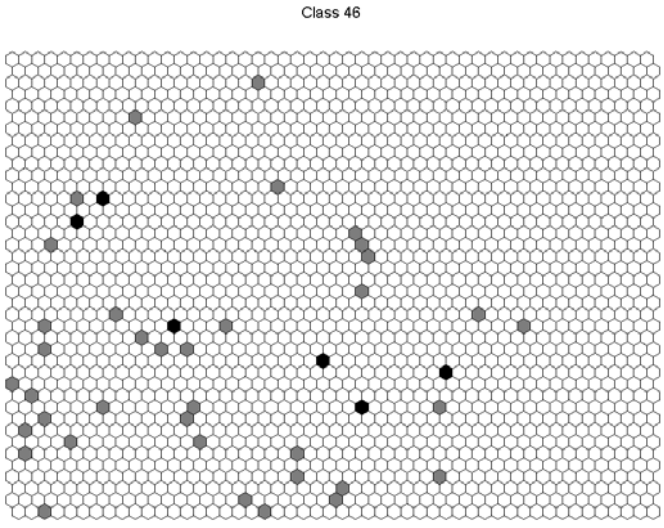


Fig. 71. The hit diagram of those 49 general prepositions that occurred 10 to 100 times in the MCRC.

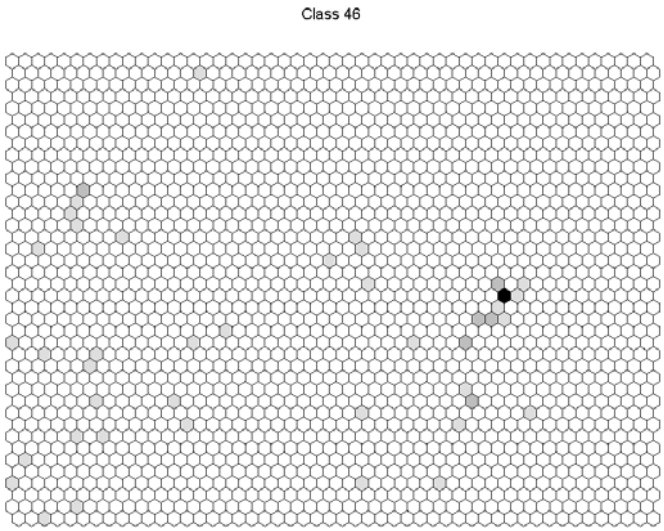


Fig. 72. The hit diagram of those 56 general prepositions that occurred 100 to 1000 times in the MCRC.

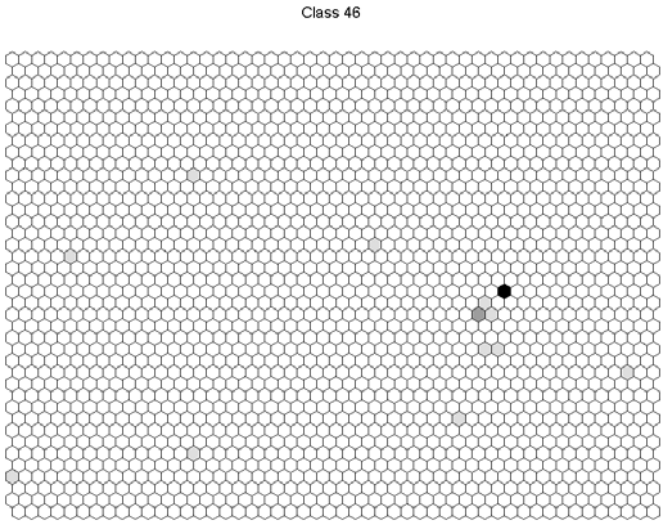


Fig. 73. The hit diagram of those 22 general prepositions that occurred 1000 to 10000 times in the MCRC.

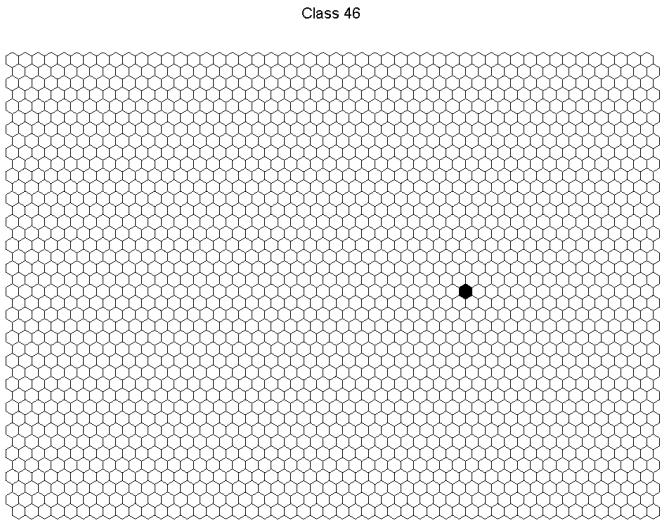


Fig. 74. The hit diagram of one general preposition that occurred at least 10000 times in the MCRC.

2.14 General adverbs (dr)

The adverbs (‘ad-verb’) are located in the hit diagrams along the main border between the nouns and the verbs. They have only few clusters in fixed places, showing that there are only few main types of adverbs, and their contexts depend very little on word frequency.

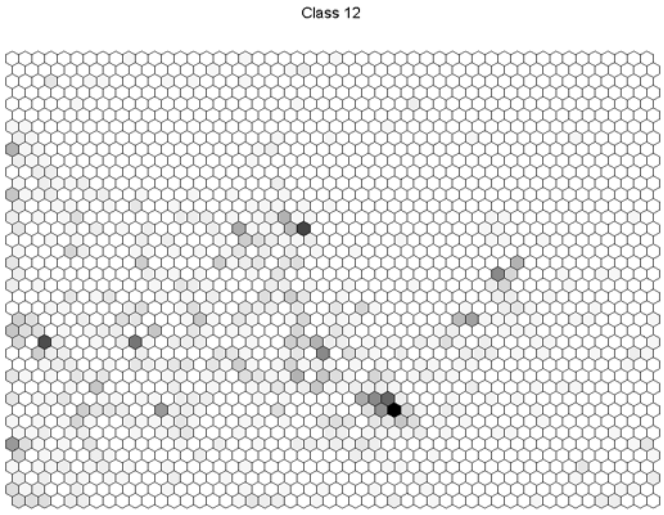


Fig. 75. The hit diagram of those 1193 general adverbs that occurred in the MCRC.

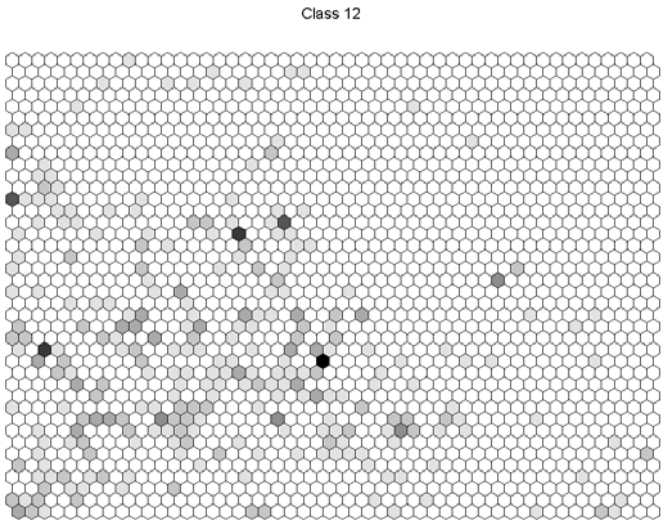


Fig. 76. The hit diagram of those 413 general adverbs that occurred 1 to 10 times in the MCRC.

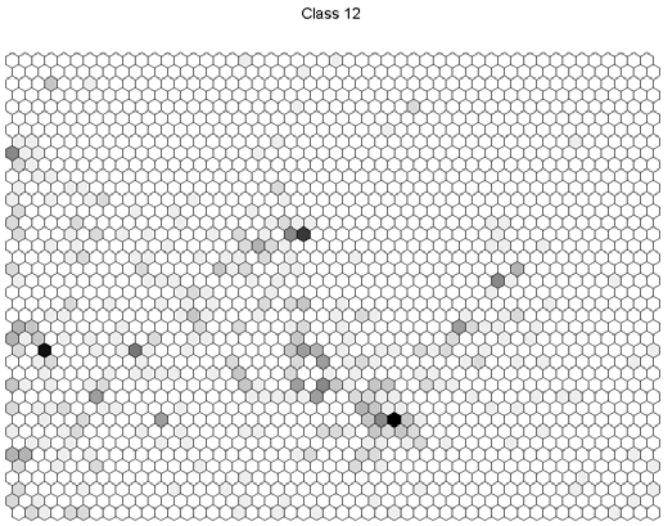


Fig. 77. The hit diagram of those 524 general adverbs that occurred 10 to 100 times in the MCRC.

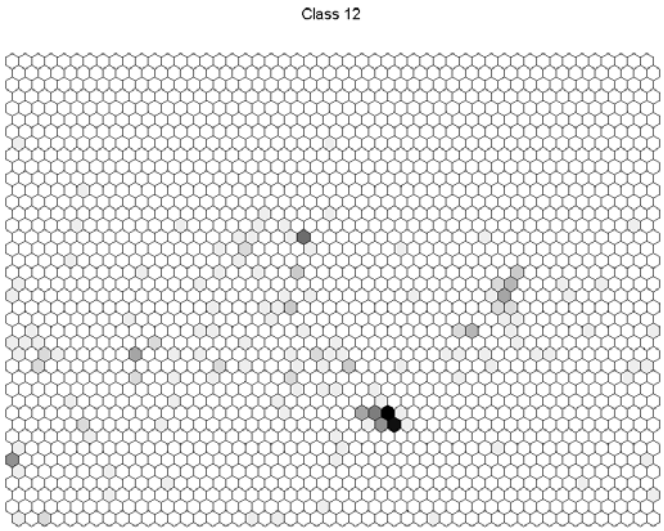


Fig. 78. The hit diagram of those 223 general adverbs that occurred 100 to 1000 times in the MCRC.

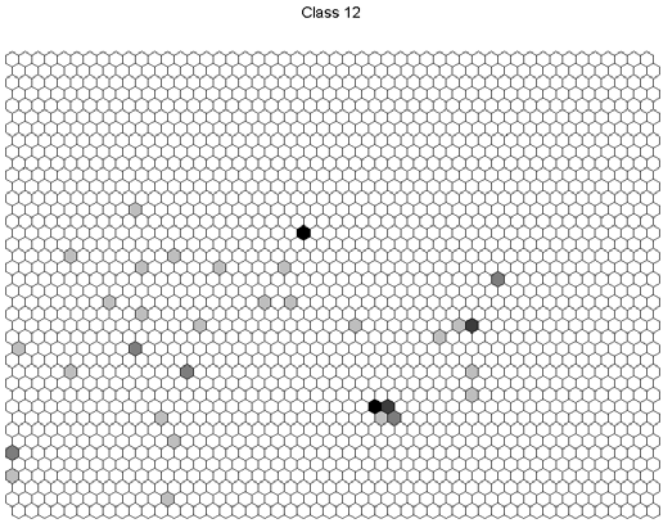


Fig. 79. The hit diagram of those 47 general adverbs that occurred 1000 to 10000 times in the MCRC.

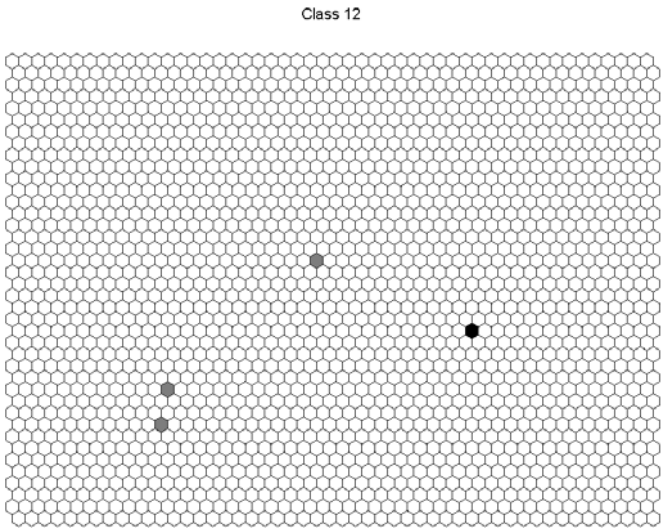


Fig.80. The hit diagram of those five general adverbs that occurred at least 10000 times in the MCRC.

2.15 Conjunctions (c)

This class is randomly scattered and not clustered, except for an empty zone at the top of the map, showing that its words do not have any clear contextual role. This could be expected, since the conjunctions occur in fixed places in the sentences and do not correlate with the rest of the words.

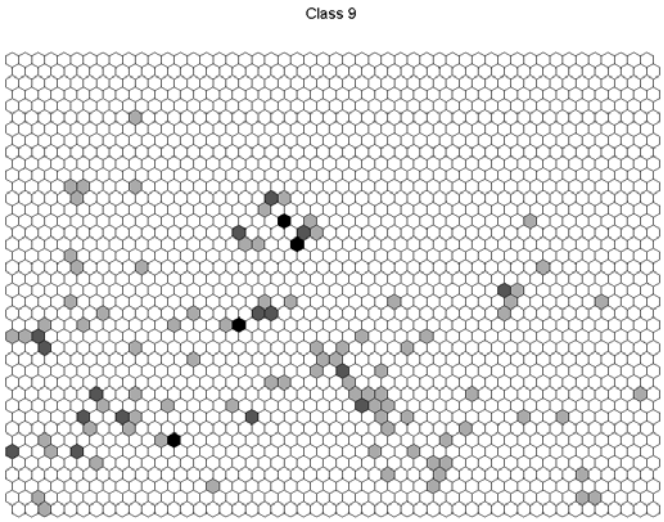


Fig.81. The hit diagram of those 128 conjunctions that occurred in the MCRC.

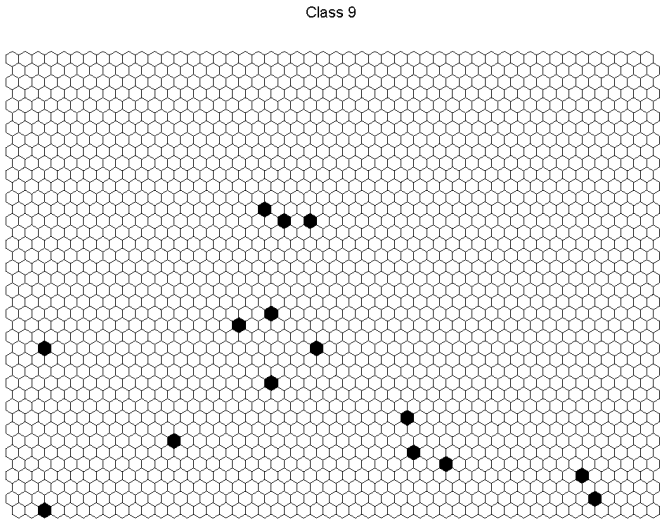


Fig.82. The hit diagram of those 15 conjunctions that occurred 1 to 10 times in the MCRC.

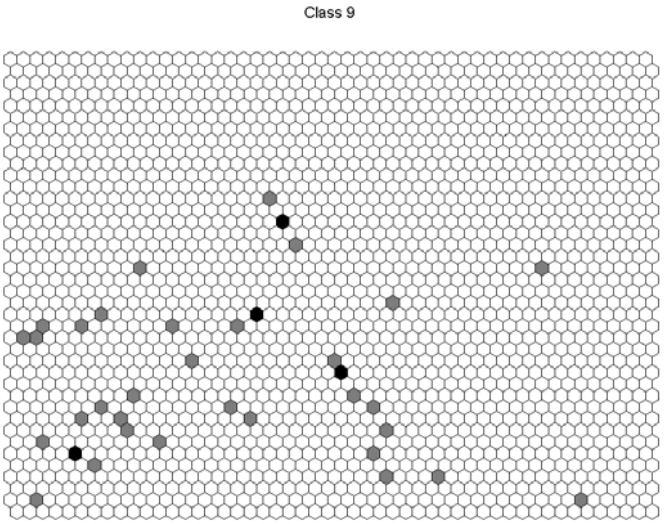


Fig.83. The hit diagram of those 40 conjunctions that occurred 10 to 100 times in the MCRC.

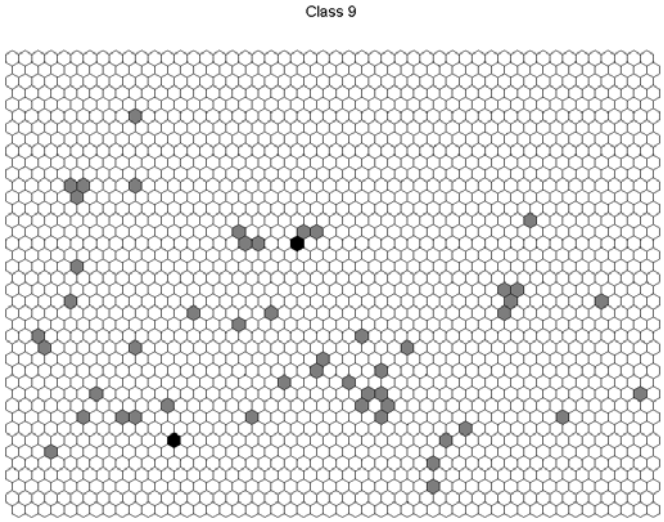


Fig.84. The hit diagram of those 53 conjunctions that occurred 100 to 1000 times in the MCRC.

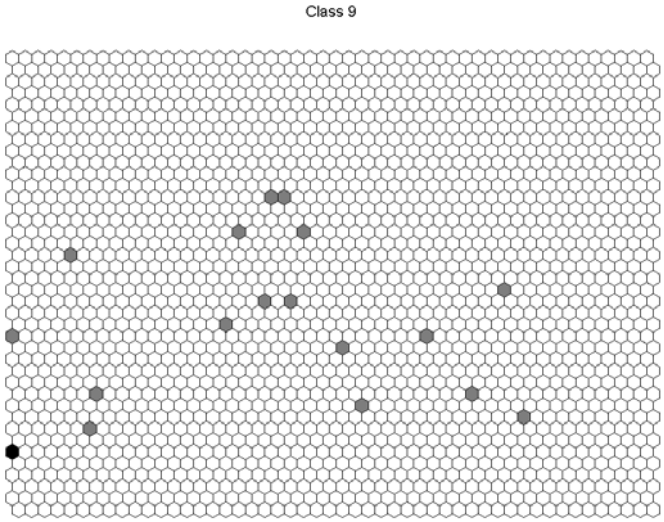


Fig.85. The hit diagram of those 19 conjunctions that occurred 1000 to 10000 times in the MCRC.

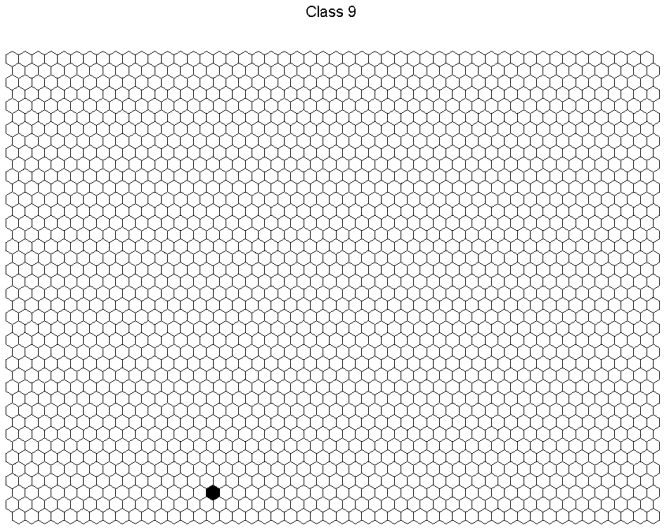


Fig.86. The hit diagram of one conjunction that occurred at lest 10000 times in the MCRC.

3 Discussion

In the previous study on contextual SOMs of Chinese words (Kohonen, 2010), the hit diagrams of some classes of words, e.g., those of the general nouns and general adjectives looked rather diffuse at the first sight. The simplest explanation may be that the sets of words of any class in general consist of several smaller subsets with a more specific semantics, and the subsets are clustered more sharply. However, such detailed information about these subsets was not available with this corpus.

The new result obtained in this study is that the hit diagrams also depend on the frequencies of the words selected for testing. In some cases, e.g., for nouns and adjectives (the diagrams of which were the most diffuse ones) this dependence is strong, whereas for some other word classes like the numerals and general adverbs it is weak.

The most natural explanation of the frequency dependence that comes into mind is that the MCRC corpus used in this work is very heterogeneous. It contains texts from newspapers, novels, magazines, TV shows, folktales, etc., written by different people. The vocabularies of the different parts, especially the sets of nouns and adjectives in them have probably very different word frequencies. Conversely, when the word frequencies during testing are restricted to certain intervals, they correlate closest with certain parts of the corpus, and then the corresponding hit diagrams look different. It would be very interesting to compare the present results with those produced by one author only and dealing with a well-defined topic area, preferably written in a traditional style.

On the other hand, it is also thinkable that the contexts in which especially the nouns and adjectives are used have changed with time, and frequent usage accelerates this semantic transformation. One fact that supports this assumption is that a hit diagram as a function of word frequency often changes gradually in the same direction (cf. Figs.29 through 32).

The semantics of the numerals seems to be most independent of frequency. There are also word classes such as the conjunctions, the hit diagrams of which look completely random and are not clustered at all. This is understandable, because the conjunctions do not correlate with the other words, and they occur always in a particular position in the sentences; so their position on the SOM is haphazard.

It is interesting that the hit diagrams of the verbs followed by nouns are well clustered in the middle of the area where the nouns have an empty space. This may be due to the almost symmetric role of these verbs with respect to the nouns.

It may be necessary to remind that the self-organizing process of the SOM can converge into alternative orderings of the mapping, all of which represent local optima. Nonetheless the mapping can be regarded as topologically correct, if the geometric relations between the clusters in the hit diagrams are the same as in the source data. The maps published in this work are samples of numerous maps that had different parameters and sizes, and have been produced under different training conditions. Nonetheless the SOM selected to this work is believed to be typical of the majority of the maps.

In the contextual SOM, the selection of the random-vector representations for the words may have an effect on the exact form of the SOM, due to statistical variations in the matching of the random vectors. These statistical variations could be eliminated for the most part if one were able to use representation vectors with extremely high dimensionalities, for which supercomputers would be needed.

One particular notice may be necessary, when looking at the graphics of the hit diagrams. In order to be able to compare histograms that contain very different total numbers of hits, one usually normalizes the intensities of the images. The MATLAB graphics does this automatically, unless other options are specified. Then, however, the clusters in low-intensity images are overemphasized and should not be compared directly with clusters in the other images. The images with very few hits such as Figs. 6, 12, 17, 22, etc. may also be misleading, since if there are only dots corresponding to a very small number of hits, most of them are displayed very black, and it seems as if there were a cluster concentrated at each of them. In order to gain a correct view of the cluster structures, one should concentrate on figures in which a great number of lexical words are included.

The main message of the work in presentation is that the word frequencies probably have an important role in all of the contextual-SOM experiments and should be taken into account especially when picking up words from the lexica for testing, preferably by uniform probability.

References

1. T. Kohonen, *Contextually Self-Organized Maps of Chinese Words*, TKK Reports in Information and Computer Science, TKK-ICS-R30 (Aalto University School of Science and Technology, Espoo, Finland, 2010).
2. H. Ritter and T. Kohonen, "Self-organizing semantic maps," *Biol.Cyb*., 61, 241-254, 1989.
3. H. L. Sun, D. J. Sun, J. P. Huang, D. J, Li, and H. B. Xing, "Corpus for modern Chinese research," in *Studies in the Chinese language and characters in the era of computers*, ed. by Z. S. Luo and Y. L. Yuan, pp. 283-294 (Tsinghua University Press, Beijing, China, 1996).

(Submitted on August 2, 2010)

TKK REPORTS IN INFORMATION AND COMPUTER SCIENCE

TKK-ICS-R25 Jani Lampinen, Sami Liedes, Kari Kähkönen, Janne Kauttio, Keijo Heljanko
Interface Specification Methods for Software Components. December 2009.

TKK-ICS-R26 Kari Kähkönen
Automated Test Generation for Software Components. December 2009.

TKK-ICS-R27 Antti Ajanki, Mark Billinghurst, Melih Kandemir, Samuel Kaski, Markus Koskela, Mikko Kurimo, Jorma Laaksonen, Kai Puolamäki, Timo Tossavainen
Ubiquitous Contextual Information Access with Proactive Retrieval and Augmentation. December 2009.

TKK-ICS-R28 Juho Frits
Model Checking Embedded Control Software. March 2010.

TKK-ICS-R29 Miki Sirola, Jaakko Talonen, Jukka Parviainen, Golan Lampi
Decision Support with Data-Analysis Methods in a Nuclear Power Plant. March 2010.

TKK-ICS-R30 Teuvo Kohonen
Contextually Self-Organized Maps of Chinese Words. April 2010.

TKK-ICS-R31 Jefrey Lijffijt, Panagiotis Papapetrou, Niko Vuokko, Kai Puolamäki
The smallest set of constraints that explains the data: a randomization approach. May 2010.

TKK-ICS-R32 Tero Laitinen
Extending SAT Solver With Parity Constraints. June 2010.

TKK-ICS-R33 Antti Sorjamaa, Amaury Lendasse
Fast Missing Value Imputation using Ensemble of SOMs. June 2010.

TKK-ICS-R34 Yoan Miche, Patrick Bas, Amaury Lendasse
Using Multiple Re-embeddings for Quantitative Steganalysis and Image Reliability Estimation. June 2010.

ISBN 978-952-60-3300-6 (Print)
ISBN 978-952-60-3301-3 (Online)
ISSN 1797-5034 (Print)
ISSN 1797-5042 (Online)