Towards a mathematical model of word class clusterings

Nordhoff, S.

Publication date
2008

Published in
Linguistics in Amsterdam

Towards a mathematical model of word class clusterings

Sebastian Nordhoff

Universiteit van Amsterdam

Croft (2001) argues that distributional analysis of word classes is doomed to failure because there is no way to know when to stop splitting word classes into subclasses. This paper discusses mathematical clustering algorithms and shows that contrary to Croft's assumption there exist hard and fast criteria to know when to stop splitting. The method exposed is applied to a subset of English lexemes first proposed by Crystal (1967). Finally, the clustering properties of typologically diverse languages are discussed in the light of the clustering model and checked against current theories of parts-of-speech. The paper concludes by affirming that clusterings can be established for any language but cannot be equated with the classical notion of parts-of-speech.

Keywords: cluster analysis, parts-of-speech, typology, categorization, dendrogram

1 Introduction

Establishing word classes with rigid boundaries on the basis of distributional analysis has a long tradition in the field, culminating in American Structuralism whose creed this procedure eventually became (cf. Aarts 2006: 369f.). Their influence is still strong today, mainly in descriptive linguistics, where discovery procedures based on distributional analysis are the mainstay of researchers who go out to the field to collect new data on undescribed languages.

These discovery procedures based on analogy met with scepticism by researchers mainly interested in anomaly (cf. Seuren 1998). Classical distributional analysis presupposes that there will be neat Aristotelian categories, and that for every lexeme, its distribution will clearly indicate to what class it belongs. From the 1960s onwards, there was rising discontent with Aristotelian categories as the boundaries between for instance noun and verb were shown to be less rigid and more fuzzy than previously thought (Ross 1972, Clark & Clark 1979). Rosch's work on semantic prototypes (Rosch 1975) dealt another blow to Aristotelian categories, albeit on a different level of analysis. Gross's (1979) work also cast doubt on the value of distributional analysis. He showed that out of 12000 French

verbs, there are none that have an identical distribution. This means that on very thorough analysis, one-member classes must be assumed if one wants to use word classes to carry information about the morphosyntactic behaviour of their members.

With distributional analysis having lost much of its credit, alternative methods for establishing word classes were proposed. Hopper & Thompson (1984) argued that word classes are a pure epiphenomenon of discourse. The Amsterdam school of Functional Grammar would not do away with the whole of morphology as a means for establishing word classes (derivation is still important in that theory), but the foundation on which the word classes are grounded was propositional function and not distribution. Head of predicate, head of term, modifier of term and modifier of predicate were the propositional functions that were determined, and lexemes were assigned to classes reflecting the possibility to be used in these functions "without further measures being taken"(Hengeveld et al. 2004, Hengeveld 1992b).

Feeling that grounding word classes on only one of discourse function, semantics or morphosyntax was too limited, a prototype approach which comprises elements from all three levels of analysis was advocated by Givón (1984) and Croft (1991). Croft (2001: 78) especially argues that morphosyntax alone cannot do the job because 'there is no way to stop splitting', following Gross.

This can be reformulated as the question: 'How many word classes can be found in my language data?', and indeed this question has given rise to heated debates over the amount of word classes in a given language (e.g. Sasse (1993), Sasse (1988) vs. Mithun (2000), Evans & Osada (2005b) and replies in the same volume.)

Questions like the above are the area of expertise of clustering methods (Halkidi et al. 2001: 111). Clustering methods can provide answers to the amount of clusters in a given data set, the dispersion of data within the cluster, the neatness of boundaries and the goodness-of-fit of clusters for a data set.

This paper argues that while it is true that on a microscopic level of analysis, distributional analysis leads to a myriad of single-item classes, these classes can nevertheless be mathematically clustered. Clustering then leads to higher-order classes with more members, and finally a single large class, which comprises the entirety of all lexemes. By putting mathematical constraints on the dispersion of clusters, the question 'how many word classes?' can be answered on morphosyntactic grounds alone.[1]

1 This does not discount the usefulness of establishing similar clusters on semantic or pragmatic grounds. Such a procedure is indeed explicitly advocated by Sasse (1993) and implicitly

The aim of this article is to outline a methodology that has proven fruitful in many other scientific domains and to propose an application of this methodology within the field of linguistics to a wider audience.[2] We will first define the concept of cluster in general terms and discuss key properties before we explain how language data can be represented for clustering purposes. We will then apply our methodology to a feature-value-matrix proposed by Crystal (1967). This matrix was intended to show the difficulties of categorization, but we will see that a cluster analysis yields good categorization even of this data. The paper finishes by squaring the results with some current theories of parts-of-speech.[3]

2 The cluster model of word classes

What is a cluster? Guha et al. (1998) define clustering as follows:

"Clustering [...] is about partitioning a given data set into groups (clusters), such that the data points in a cluster are more similar to each other than points in different clusters."

The important notion here is that there is no requirement for identity. Being very similar is enough for two points to end up in the same cluster. Clusters are found in many different scientific disciplines. One area that is particularly interested in good ways to establish clusters are the social sciences, particularly customer research. But other fields have made use of cluster analyzes as well (Han & Kamber 2001: 336), for instance image segmentation[4], genetics[5] or geography[6]. In these disciplines, mathematical methods have been developed to identify clusters in a given data set. Such a data set can be information about age, sex, income and shopping habits of a customer, about presence or absence of mutations in certain

---

present in Hengeveld et al. (2004), Hengeveld (1992b) and maybe Croft (2001).

2 As such, discussions of primarily computational implications of the methodology proposed are reduced to a minimum in this paper. Considerations of efficiency in runtime, memory, robustness to noise and behaviour of certain algorithms in borderline situations as well as a discussion of which algorithms would best suit a linguist's needs are deliberately left out in order to provide maximum accessibility also to people outside the field of computer science.

3 One anonymous reviewer remarks that the applicability of this method is not limited to parts-of-speech and can indeed be extended "to any other subdomain of linguistics where categorization is an issue". This is indeed possible, syntactic categories or grammatical relations come to mind, not to mention semantics. Parts-of-speech are a prime testing ground for this method because of the overwhelming importance of categorization for this field, which may have seen more heated debates about the 'right' classification of items than other fields of linguistics.

4 e.g Pal & Pal (1993).

5 e.g. Eisen et al. (1998).

6 e.g. Guo et al. (2003).

genes, but it can also be information about the morphosyntactic properties of a lexeme. A transfer of these methods developed in the social sciences and the natural sciences to the domain of linguistics might help to clear some persistent categorization problems.[7]

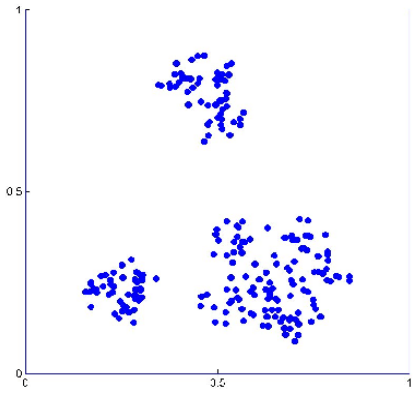


Illustration 1: *Clusters in 2-dimensional space*

Figure 1 shows three clusters in 2-dimensional space. Humans immediately perceive three clusters as Gestalt. For computers, this is a bit more difficult. We have to define a cluster in a mathematical way in order to process them computationally. What is a cluster mathematically speaking? A cluster is a set of vectors that share the property of being more similar among themselves than they are similar to other vectors. This similarity can be computed by a *distance metric*. The simplest cluster is the atomic cluster, which consists only of one element. This element is of course maximally similar to itself. The distance to itself is 0 . Bigger clusters are formed by comparing the distances between vectors and grouping together vectors with small distances.

Non-zero distance does not preclude that two vectors end up in the same

7 In the domain of linguistics, hierarchical cluster analysis has been known for quite some time in research on part-of-speech taggers (cf. Ushioda 1996, Wang & Vergyri 2006). These are programs that automatically assign parts-of-speech to words in an electronic corpus. Most of research in this domain has been done on English, a language whose parts-of-speech system is comparatively well described. Porting the findings to typology is however difficult since for most languages we do not have suitable annotated electronic corpora. Also, parts-of-speech taggers already rely on some assumptions about the parts-of-speech system of the language (but see Schone & Jurafsky 2001). As the focus of this paper is to determine methods to describe the system in the first place, methods that already rely on a description are not suitable.

cluster, as is apparent from Figure 1. Vectors in one cluster can be distant from one another, provided that these differences are smaller inside the cluster than to elements outside the cluster. We can speak of *intra-cluster variance*, which should be low, and *inter-cluster variance*, which should be high.

A low intra-cluster variance indicates a *dense* cluster. This means that good predictions can be made about the behaviour of the members of that cluster. It is very *informative*. A high intra-cluster variance indicates a *dispersed* cluster. Here, fewer predictions can be made.

A high inter-cluster variance indicates that the vectors of the cluster are very different from other clusters. The separation between the clusters is neat. A low inter-cluster variance indicates that the vectors of the cluster are not very different from other clusters. This means that the boundary between the two clusters is not very neat. In linguistics, so called 'squishes' (Ross 1972) for instance would be clusters with low inter-cluster variance.

Let us discuss the concepts in italics in the above paragraph in turn.

2.1 A lexeme as a vector

Clustering methods treat the items they cluster as vectors n-dimensional space.[8] We must therefore find a way to represent the grammatical properties of a lexeme as a vector. This can be done by means of feature-value-pairs.[9]

Particular grammatical properties of a token (e.g beans) are often described by feature-value-pairs, for instance number:pl. In this example, the feature number of the lexeme bean has the value plural.

When talking about lexemes (types), feature-value-pairs can also be used. A particular feature of a lexeme is for instance the possibility to attach a morphological plural. Another feature would be the possibility to attach a marker for past-tense.

(1) *bean-s *bean-ed*

We can say that the lexeme beans has the value yes for the feature can be combined with the plural marker -s and has the value no for the feature can be combined with the past-tense marker -ed.

(2) a. *bean*: MORPHOLOGICAL PLURAL= +

8 Readers familiar with linear algebra may want to skip to section 3.1.

9 Another method would be co-occurrence (Wang & Vergyri 2006).

b.*bean*: MORPHOLOGICAL PAST TENSE= -

Different lexemes can have different values for these features.

(3) a. *communicate*: MORPHOLOGICAL PLURAL= -

b. *communicate*: MORPHOLOGICAL PAST TENSE= +

(4) a. *saddle*: MORPHOLOGICAL PLURAL= +

b. *saddle*: MORPHOLOGICAL PAST TENSE= +

(5) a. *happy*: MORPHOLOGICAL PLURAL=-

b.*happy*: MORPHOLOGICAL PAST TENSE= -

These feature-value-pairs can be represented in a table as in Figure 2 below. Because we will use them for computational operations later, + is replaced by 1 and - is replaced by 0 . We can now say that bean has the value 1 for the dimension morphological plural.

	bean	communicate	saddle	happy
morphological plural	1	0	1	0
morphological past tense	0	1	0	0

Illustration 2: A simple table of features and values

Each feature can be seen as a dimension in vector space. A lexeme's value in a given dimension can be 1 or 0 in our initial model. We agree that the dimension morphological plural shall always be in the first row, and the dimension morphological past tense always in the second row. We furthermore agree that the nth component of a vector can be indicated by an index, $bean_1 = 1$, $bean_2 = 0$.

(6) $bean = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ $communicate = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ $saddle = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ $happy = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

These vectors only have two dimensions and can thus be plotted on paper (Figure 3). Later, we will treat vectors of higher dimensions, which are more difficult to visualize.

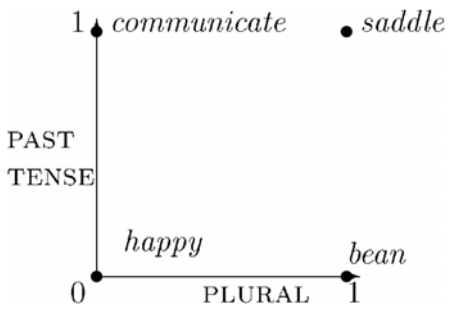


Illustration 3: A very simple two-dimensional feature space with four lexemes.

2.2 Distance metrics

2.2.1 Distances between lexemes and vectors in 2-dimensional space

How can distances between vectors be computed? In 2-dimensional space, we can use the Pythagorean theorem (7).

(7) $d(a, b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$

(8) a. $d(bean, bean) = \sqrt{0^2 + 0^2} = \sqrt{0} = 0$

b. $d(bean, happy) = \sqrt{1^2 + 0^2} = \sqrt{1} = 1$

c. $d(bean, saddle) = \sqrt{0^2 + 1^2} = \sqrt{1} = 1$

d.. $d(bean, communicate) = \sqrt{1^2 + 1^2} = \sqrt{2} \approx 1.41$

We see that *bean* and *communicate* are more distant than *bean* and *happy* or *bean* and *saddle*. This mirrors our non-mathematical intuition that nouns and verbs are quite distant from one another, but adjectives are somewhere in between, with less distance separating them from nouns, and from verbs. *Saddle* as a lexeme which can function as a noun and a verb is also less distant from the lexemes representing the classical categories of nouns and verbs here.[10]

10 In the clustering model discussed here, it is not possible for a vector to belong to more than one cluster. Furthermore, semantic differences are disregarded, only the distributional behaviour of the lemma is analyzed. We see in Figure 4 that the English lexemes which are found in both the traditional classes of nouns and verbs, such as *saddle*, show a considerable distance from both canonical verbs and canonical nouns. This means that they will not cluster easily with either of them, but remain a separate cluster (depending on Θ and τ, see below). These facts then need to be interpreted by the linguist. One interpretation, the 'classical' one,

If we compute all the distances between the four lexemes given above, we can arrange them in a distance matrix (Figure 4). The distance metric shows us how similar the vectors that represent the lexemes are. The shorter the distance the more similar, the longer the distance the more different. This distance metric is for the entirety of all vectors under discussion, but distance matrices for subsets can also be computed by only considering a subset of vectors.

	bean	happy	saddle	communicate
bean	0	1	1	1.41
happy	1	0	1.41	1
saddle	1	1.41	0	1
communicate	1.41	1	1	0

Illustration 4: A very simple distance matrix

2.2.2 Distance between lexemes and vectors in n -dimensional space

We see that in two-dimensional feature space, we can compute the distance between two vectors using Pythagoras. But real-word lexemes have much more features than the two used for this example. Linguists have unearthed many more dimensions in which lexemes differ: gender, aspect, negation, subcategorization patterns, passivization, etc.

We have to leave 2-dimensional space and go to *n*-dimensional space, where *n* is the number of features we wish to investigate. Remember that every feature represents a dimension. We will first illustrate the method on a hypothetical example before we apply it to a real-world matrix in section 5. Suppose we have investigated 64 grammatical features which we want to use to characterize our lexemes.[11] The vector for a lexeme bean could then be something like (9), which is a notational variant of the column notation used above in order to save space. A 1 indicates that the feature can grammatically be expressed on the lexeme, while a 0 indicates that this is not the case.

(9) **bean**=(0111000101011100100111001001011100010101110010011100100101010 1)

In the 2-dimensional space, we used Pythagoras to calculate the distance between two points. A similar formula exists for any n-dimensional space.

(10) $d(a,b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + (a_3 - b_3)^2 + \ldots + (a_n - b_n)^2}$

would then be to analyze this clustering as a reflex of double class-membership. Another interpretation would be that there are four word-classes in the sample, N, V, ADJ and N+V. Which of these interpretations is to be preferred is independent of the clustering as such, and thus this issue will not be discussed any further here.

11 Which features these are precisely is not important for the following discussion.

This formula is used to calculate the Euclidean Distance.[12] If we have established the vector representation for every lexeme, we can calculate the distance matrix (Figure 5).

3 Clustering in vector space

3.1 Method

How does the clustering procedure work? One of the easiest ways of clustering is the AGNES algorithm (Han & Kamber 2001: 355). We take a distance matrix and find the two vectors with the lowest distance. These two vectors form the first cluster then. They are merged, i.e. their orthographical forms are written together between braces, and their centroid is assumed as their new common vector. This means we withdraw two vectors and add one new one, the one that represents the cluster. We then compute a new distance matrix and iterate the procedure until only one vector is left.[13]

3.2 Measures of dispersion and information loss

We can calculate the dispersion D of a cluster based on its member vectors. The simplest method is to compute the maximum distance between any two points in the cluster. Other measures are variance, standard deviation or mean deviation.[14] We can set a threshold Θ for the maximum dispersion we are willing to tolerate within a cluster.

The difference between the dispersion in a given cluster and the dispersion in the next highest supercluster tells us about the quality of this clustering step. It symbolizes information loss when we merge two subclusters into a supercluster, or information gain when we split a supercluster into two subclusters. We will

12 For other distance metrics see Theodoridis & Koutroumbas (2006: 358f.). All mathematical terms in italics without reference are explained in any good introduction to linear algebra and statistics.

13 The algorithm described here makes use of agglomerative hierarchical clustering. There exist many other approaches to clustering that all have their particular advantages and drawbacks, whose discussion is clearly outside the scope of this paper. The interested reader is referred to Halkidi et al. (2001) for a brief overview or to Jain & Dubes (1988), Han & Kamber (2001) or Theodoridis & Koutroumbas (2006) for a more thorough introduction. Guha et al. (1999) propose the ROCK algorithm, which seems the most promising for linguistic data.

14 See Sharma (1996) for more metrics, combination of metrics, discussion and evaluation.

symbolize it with ΔD.[15]

(11) $D_{supercluster\{a,b\}} = D(a, b) - (D(a) + D(b))/2$

A low value for ΔD indicates that there is little difference in the information contained in the two subclusters as compared to the supercluster, while a high value indicates that they differ considerably.

We can set a threshold τ for the information loss we are willing to tolerate or the information gain we require.

	happy	sad	comm.	rejoice	apple	bean	hammer	saddle
happy	0							
sad	1.00	0						
communicate	5.38	5.47	0					
rejoice	5.56	5.65	1.41	0				
apple	6.00	5.91	7.68	7.81	0			
bean	5.74	5.65	7.87	8.00	1.73	0		
hammer	7.68	7.74	6.16	6.00	5.00	5.29	0	
saddle	7.93	8.00	5.83	5.65	5.38	5.65	2.00	0

Illustration 5: A distance matrix for 8 lexemes and 64 dimensions. Low values that will be merged soon are in boldface.

If we take the data from Figure 5 we can illustrate the clustering procedure. First we establish the lowest distance between two different lexemes. This is *d(happy, sad)* = 1.00. We then compute the new center of {*happy,sad*}, suppose it is

(12) c(h,s) =(0 1 1 1 0 0 0 1 0 1 0 1 1 1 0 0 1 0 0 1 1 1 0 0 1 0 0 1 0 1 1 1 0 0 0 1 0 1 0 1 1 1 0 0 1 0 0 1 1 1 0 0 1 0 0.5 0 1 0 1 0 1 0 1)

In the dimension that is underlined, the new vector differs from the old vectors. We recalculate a new distance matrix and establish the pairing with the lowest distance. {*happy, sad*} now counts as one vector for the calculation of the distance. Hence we have only 7 vectors left. This is continued until only one vector is left. We can present the result of the hierarchical clustering as a dendrogram (Figure 6). On the *x*-axis, we see how much dispersion a given cluster has (intra-cluster variance). The horizontal branches between two clusterings represent the information loss/gain between a subcluster and a supercluster. It also describes the inter-cluster variance between a subcluster and its sister clusters.

As an example, the singleton cluster {*happy*} has a dispersion of 0. Its

15 Again, see Sharma (1996) for more sophisticated calculations of ΔD.

supercluster has a dispersion of 1.00. We can calculate the information loss $\Delta D\{happy,sad\}$ we incur by grouping

happy together with sad as $\Delta D\{happy,sad\} = 1 - (0 + 0)/2 = 1$. For the next higher clustering we get:

(13) $D_{supercluster\{a,b\}} = D(h,s,r,c) - (D(h,s) + D(c,r))/2 = 9.8 - (1 + 1.4)/2 = 8.6$

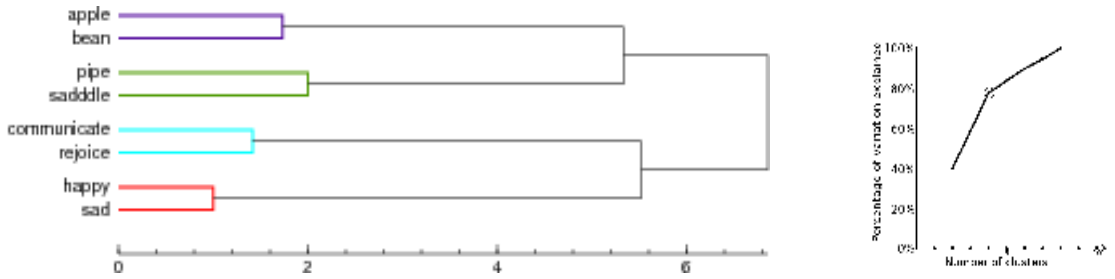


Illustration 6: A clustering of 8 lexemes in 64 dimensions

3.4 Cutting the dendrogram and establishing major word classes

The dendrogram as such only presents one cluster. We have to cut it to get more clusters. We can establish a threshold Θ for the maximum dispersion we are willing to tolerate for a cluster. At that threshold, we cut the dendrogram (cf. Theodoridis & Koutroumbas 2006: 407). In Figure 6, this is shown for Θ =9.0.[16] Instead of setting a threshold Θ for the absolute dispersion, we can also set the threshold τ for minimum difference in dispersion, i.e. minimum information gain. Let τ = 1.9. In that case the four long branches in the dendrogram would be cut, but additionally, the cluster {*saddle,hammer*} would not meet the criterion with $\Delta D\{saddle, hammer\} = 2.0 > 1.9$. {*saddle*} and {*hammer*} would form two singleton clusters then.

A combination of the two thresholds is also possible. In that case, Θ can be used to determine major word classes, while τ can be used to see whether these should be split into minor word classes. The number of word classes is then a function of the values of Θ and τ and the distance matrix.

(14) $NWC = f(M, \Theta, \tau)$

NWC is thus not invariant for a given language, but depends on the features chosen, which influence M, and the values chosen for Θ and τ .

16 See Boberg & Salikowski (1993) for a better algorithm and a more thorough discussion.

Finally, there are also methods to compute the amount of variation explained by a partition of a data set into clusters (Ray & Turi 1999). This can be done for all values *0* ... *n*. The result is a graph as shown in Figure 6b. A higher number of clusters always explains more variation, but there is normally a drop in additional information provided after the first clusterings. This is reflected in a 'knee' in the plotted graph, in our example for n=4.[17]

By keeping the values for Θ and τ constant, we have a hard and fast criterion to know 'when to stop splitting', contra Croft. The objection that the researcher can choose any value for Θ and τ can easily be countered with the 'knee' method explained above.

4 Typology of clusters

Clusterings can differ in the neatness of their major word classes and in the information gain that subclusterings provide. Figure 7 shows some clusterings that differ in the number of major word classes and the neatness of the separation. The data for the clustering in Figure 7 are of hypothetical matrices specially constructed to highlight these types, which in turn are modeled on parts-of-speech systems described in the literature. The use of construed data sets is necessary at the macroscopic level as of now because the cluster analysis of the lexemes of even one actual language is a vast task and cannot be undertaken in this paper outlining a technique new in the field. It is hoped that cluster analysis will become more widespread in the future and that more data will become available over time. Actual data to illustrate clusterings on a microscopic level on the other hand is available and will be dealt with in the next section.

17 More information on this method can be found in Salvador & Chan (2004).

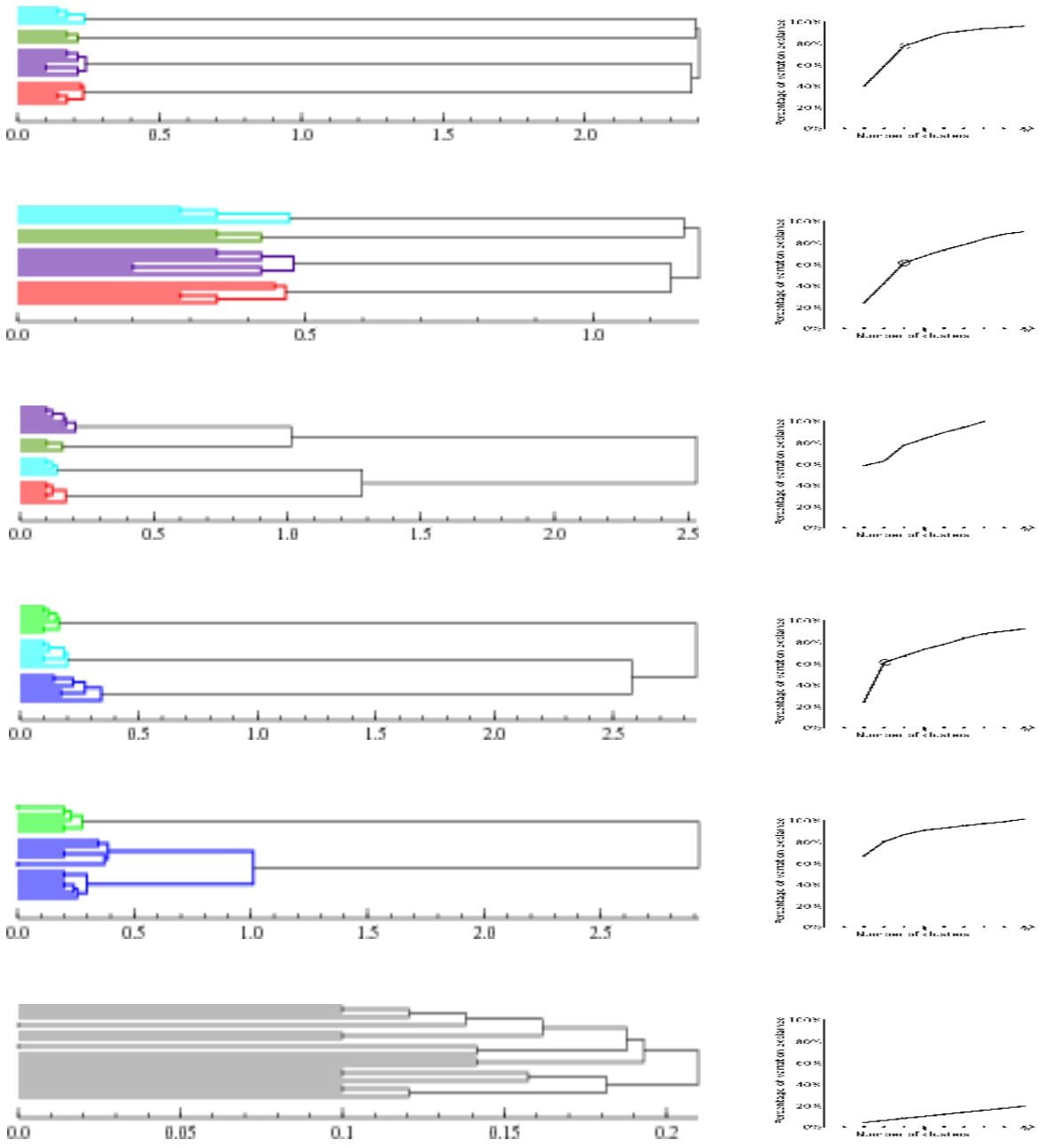


Illustration 7: Schematic clusterings for various attested PoS-systems

Looking at Figure 7a), we encounter a first division into two subclusters which provides little information gain. These two clusters are then in turn divided into 2 subclusters each, providing a huge information gain. Further subclustering yields only little information. This cluster is thus a representation of a 4-class-

system with neat boundaries, like claimed for instance for Georgian (Hengeveld et al. 2004).

Figure b) resembles a), but the information gain from 2 to 4 classes is less important. This means that the subclusters do exist but are less differentiated than in a). An example of such a type would be English.

Figure c) resembles a), but the initial split into two classes is more informative. The decision whether there are 2 or 4 classes is not straightforward. The 'knee' is bent less. Such a clustering would be predicted for Murrinh Pata (Walsh 1996), where there is a class of Nouns and a class of Verbs, but additionally there are nouny verbs called Nerbs and verby nouns called Vouns. It is not clear whether Vouns and Nerbs are on a par with Nouns and Verbs, or subordinate to them.

Figure d) shows an initial repartition into 2 clusters, which is not informative, and then a further division of one into two subclusters, yielding a total of three highly informative clusterings. Such a clustering would be expected for neat 3-class systems, like e.g. Ket (Hengeveld et al. 2004).

Figure e) also shows 3 classes, but it is not clear whether the lowest two of them should be merged or split. As such, it resembles c) Such a system can be found in languages where the category of adjective shows very similar behavior to another category (normally N or V, Bhat (1994), Wetzer (1996)).

Figure f) finally shows a system where no clear clustering can be found. Clusters can be established mathematically, but they carry very little information. Languages whose parts-of-speech system has been described as amorphous are Samoan (Mosel & Hovdhaugen 1992), Tongan (Broschart 1997), Cayuga (Sasse 1988) and Mundari (Hengeveld & Rijkhoff 2005).

While the first three clusterings are easy to interpret, this is not the case for the last three. This might be the reason why there are so many discussions over the 'right' number of word classes in a given system. The answer whether Cayuga has one word class (Sasse 1993, Sasse 1988), or two (Mithun 2000) boils down to the question whether one prefers a high value for τ , which would favor Sasse's point of view, or a low value, which would speak in Mithun's favor.[18]

A similar argument can be made for the South East Asian Adjectives (e.g. Lao, Enfield 2004). Depending on the value of Θ they will constitute a major word class or be subsumed under the Verbs. Depending on the value of τ , they will be granted the status of minor word class or not.

18 The same applies *mutatis mutandis* to Evans' (2005) critique of "monocategorialists".

The recognition of the importance of the internal structure of word classes in particular languages means that future arguments about the absolute number of word classes should be complemented by additional information about the parameters of the clustering process which produced them, namely Θ and τ . In the absence of resources to do an actual cluster analysis, researchers can state how the number of word classes is expected to vary if Θ and τ are set to low or high values. This should have little influence on clusterings like a), but much more influence on clusters like f).

5 Application

The discussion up to now has concentrated on theoretical aspects of the clustering method and the interpretation of hypothetical dendrograms. Let us now turn to a practical application to show the usefulness of this approach, based on data from Crystal (1967). Crystal analyzes some features of NPs referring to time and lists the values of these features for a number of NPs. His main aim is to show that distributional differences exist even among a semantically quite homogeneous class such as temporal nouns (Figure 8). A quick glance at this matrix does not reveal any obvious clusterings of lexemes. But Crystal already proposes to measure the differences between lexemes and compute their distance:

> the degrees of difference [...] might then be quantified in terms of the number and rank of criteria applicable and inapplicable, and these words said to be verifiably "nearer" to one class than the other [...] The problem then becomes on a par with other "higher" level problems, such as whether to take two or more clearly distinct groups of words as separate classes, or as sub-classes within one major class.

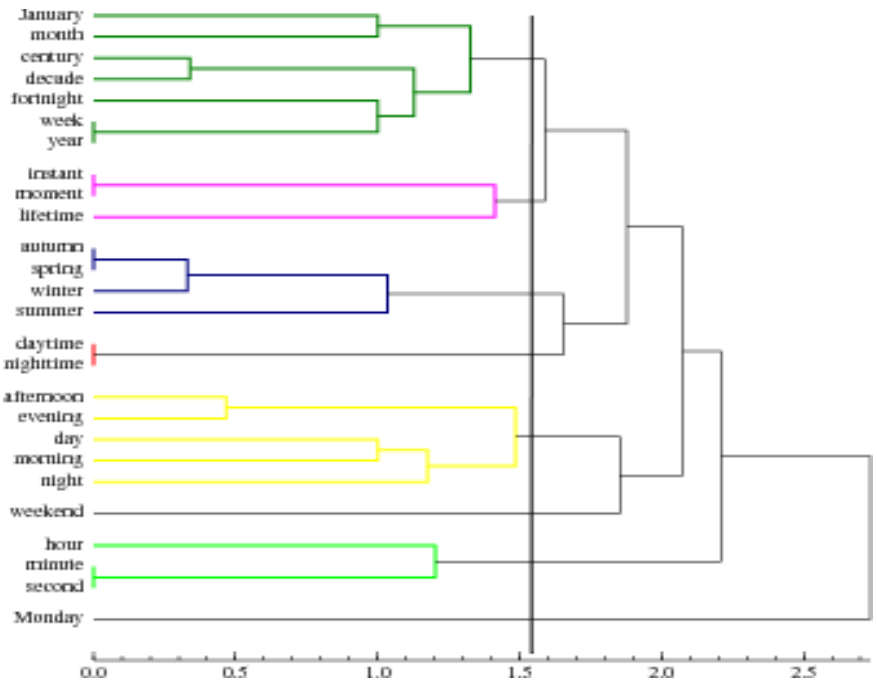
Crystal's idea is exactly the one adopted in this paper: criteria applicable and inapplicable are features with values 0 or 1, 'nearer' beacons towards distance metrics, and the identity of microscopic clustering problems with macroscopic ones is foreshadowed. In the following, we will apply the clustering technique to Crystal's own matrix of temporal nouns.

	in a N or two	in that N	in the NSG (npm)	in a N (npm)	In ø NPL (npm)	in ø NPL	ø NPL	on the NSG	on a NSG (npm)	on ø NPL (npm)	on ø NSG	at that N	at the NSG	at ø NSG
a afternoon	+	+	+	+	–	–	+	+	+	+	–	–	–	–
b evening	+	+	+	+	–	–	+	+	? +	? +	–	–	–	–
c weekend	+	+	? +	+	–	–	+	+	? +	+	–	+	+	–
d night	+	+	+	? +	–	–	? –	+	–	–	–	–	–	+
e morning	+	+	+	+	–	–	+	+	–	–	–	–	–	–
f Monday	+	–	–	–	–	–	+	+	+	+	+	–	–	–
g January	+	+	+	+	–	+	–	–	–	–	–	–	–	–
h hour	+	+	+	+	? –	–	–	+	–	–	–	+	+	–
i minute	+	+	–	+	+	–	–	+	–	–	–	+	+	–
j second	+	+	–	+	+	–	–	+	–	–	–	+	+	–
k day	+	+	+	+	–	–	–	+	–	–	–	–	–	–
l summer	? +	+	+	? –	? –	–	? +	–	–	–	–	–	–	? +
m winter	? +	+	+	? –	? –	+	? +	–	–	–	–	–	–	? +
n spring	? +	+	+	? –	–	+	–	–	–	–	–	–	–	? +
o autumn	? +	+	+	? –	–	+	–	–	–	–	–	–	–	? +
p month	+	+	+	+	+	+	–	–	–	–	–	–	–	–
q week	+	+	+	+	+	–	–	–	–	–	–	–	–	–
r year	+	+	+	+	+	–	–	–	–	–	–	–	–	–
s decade	+	+	–	+	? +	–	–	–	–	–	–	–	–	–
t century	+	+	–	+	? –	–	–	–	–	–	–	–	–	–
u fortnight	+	+	+	+	–	–	–	–	–	–	–	–	–	–
v instant	+	? +	–	+	–	–	–	–	–	–	–	+	–	–
w moment	+	? +	–	+	–	–	–	–	–	–	–	+	–	–
x lifetime	–	? +	–	+	–	–	–	–	–	–	–	–	–	–
y daytime	–	–	+	–	–	+	–	–	–	–	–	–	–	–
z nighttime	–	–	+	–	–	+	–	–	–	–	–	–	–	–

Illustration 8: Crystal's matrix. npm= no postmodifcation.

Crystal made use of graded grammaticality judgments. Our distance metric, the Euclidean Distance, can accommodate this by assigning decimal fractions to these judgments. '-' will be 0 , '?-' will be 0.33, '?+' will be 0.67, and '+' will be 1. We can now compute the distance matrix, apply the hierarchical algorithm outlined above and print the dendrogram (Figure 9).[19]

19 All trees are built with Peter Kleiber's excellent programs available at



`http://www.let.rug.nl/ kleiweg/clustering/` . Actually, for aesthetic reasons, the trees are built with the group average distance metric instead of the unweighed centroid metric outlined above. The differences are marginal.

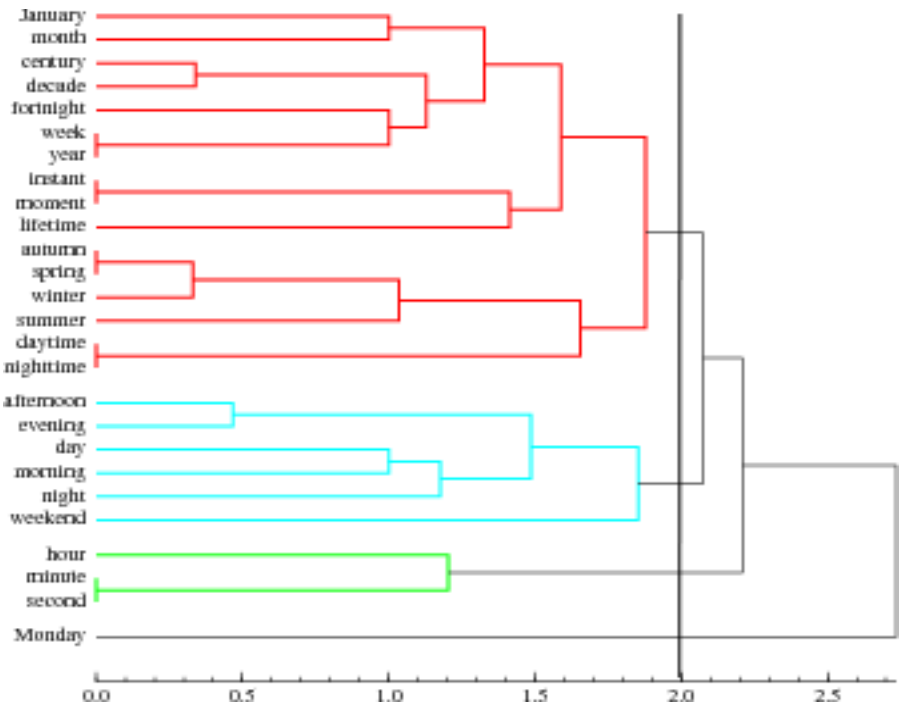


Illustration 9: Clustering Tree for Crystal's data, pruned at different distances

We can now define the threshold Θ for the maximum variation we are willing to tolerate within a cluster. Figure 9 shows two groupings for Θ=1.55, which yields 8 major groupings, and Θ=2.0, , which reduces these eight to four.

On inspecting the eight groupings, we find that morphosyntactic similarity correlates to semantic similarity. Numbering the clusters from top to bottom, we find: 1) super-diurnal expressions denoting time-spans of more than one day 2) a heterogeneous clustering 3), the seasons 4) day- and nighttime 5) subdivisions of the 24h-day 6) a singleton cluster for weekend, 7) chronometrical expressions, and finally again a singleton cluster 8), only comprising Monday.[20] A certain correlation between these morphosyntactic clusterings and semantic domains is apparent.

Allowing for more intra-cluster variation by setting Θ=2.0, we get 4 classes, one of super-hebdomadary expressions, one of sub-hebdomadary expressions, one of chronometrical expressions and one for the days of the week, again with the single member Monday. We find that our first cluster (the biggest one) also comprises instant and moment, which are not super-diurnal. This shows that the mapping between semantics and morphosyntax is not trivial.

What is the 'right' number of clusters for Crystal's matrix? Figure 9 resembles very much Figure 7f. There does not seem to be a lot of significant internal structure to the cluster. So we can say that there is only 1 cluster, just as we would assume just one word class for Samoan. A microscopic dendrogram that yields three clusters is shown in Figure 10, based on data from Quirk (1965: sec.8), who established a matrix to show a gradient transition between these verbs. Inspecting the resulting cluster, we find that the transition is less gradient than presumed.

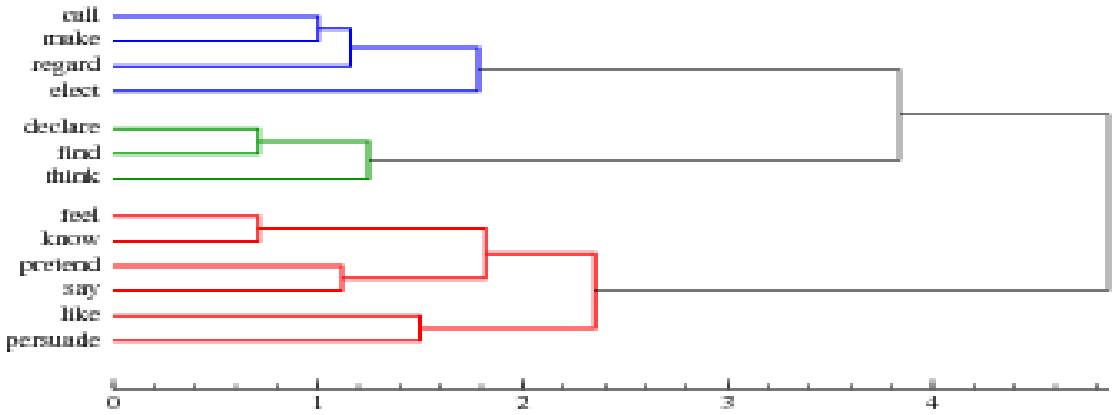


Illustration 10: dendrogram built from Quirk's matrix

20 This last cluster would probably also host the other days of the week and be a meaningful cluster, but they were not included in this sample.

We have shown that even in Crystal's or Quirk's highly diverse matrices, it is possible to establish form classes on morphosyntactic criteria alone. These can then be related to semantics.

Crystal's data are only about temporal expressions, but it is easy to conceive a broader study which takes into account more lexemes and more features. While being close to Figure 9 on the microscopic level, the literature of word-class systems of the world suggests that on a macroscopic level it would resemble one of the types in Figure 7 a-f).

Resources permitting, it is possible to analyze large sets of lexemes on a plethora of features and have them automatically clustered. The semantic interpretation of these form classes remains a task for the human linguist.

It might be objected that this clustering shown in Figures 8 and 9 is based on only 364 data points. This is very little in a time where good algorithms for treating several million data points are discussed. However, the use of such a small set is warranted by a number of reasons: 1) The aim of this section is to show micro-variation within a cluster and exemplify the methodology. Taking many more data points will not help this aim any better than the 364 data points used here. 2) This study is on the application of clustering techniques in linguistics, not on their acutal computational implementation. Discussions of computational power, runtime or footprint of a certain implementation address applicability to large data sets, but no such discussion is attempted here. Such discussions can be found in Jain & Dubes (1988), Han & Kamber (2001) and Theodoridis & Koutroumbas (2006). 3) The clustering approach advocated here intends to put parts-of-speech discussions on a more empirical basis. This methodological quest for empiricity does not entail that a clustering of all lexemes in one language has to be provided in one go, any more than theories of sociological survey methodology have to provide the answers to the polls they help to design. 4) Discussions of parts-of-speech systems in the linguistic literature usually cite some isolated examples, of which some isolated properties are then discussed. These isolated examples are taken as emblematic for their class without further justification. This paper proposes a method for avoiding ad hoc examples. As such, it can of course not give the final answer to the question of parts-of-speech. No detailed account of anything close to the majority of the lexicon of even only one language (e.g English) can be given. But even without this, I contend that the data set, provided here for exemplification purposes on a microscopic level, need not shy away from data sets that are used in the literature to base actual macroscopic claims on.

6 What do clustered word classes mean?

Can these clustered lexemes save Aristotelian categories? The answer to this is no. Strictly speaking, the clusters do not denote classes with clear boundaries but rather sets of lexemes that are more similar to each other than to other lexemes outside the set. Maurice Gross is right when he states that every lexeme is distributionally unique. But this does not have to be a reason to throw in the towel like Croft (2001) suggests. No two lexemes behaving in the same way can be found. Nevertheless, there is something to be found: lexemes that behave in very similar ways. This similarity can be computed, and depending on the interest of a particular researcher a more or less granular clustering can be chosen.

The 'meaning' of these clusters can be equated with their centroid. This vector designates the center of the cluster and thus shows how an 'average' member of that cluster behaves. Any lexeme may diverge from that centroid vector in any dimension, but the probability is that it will rather not.

Clusters thus do not provide an all-or-nothing approach, but rather are a probabilistic tool to describe the probable behaviour of a lexeme. The centroid gives the average behaviour, while the dispersion informs us about the likelihood that a given lexeme's behaviour is close to the centroid's.

This is parallel to the application of cluster analyzes in marketing. Suppose I know that a potential customer is 55 years old, well-educated, has a high income and lives in New York. In my cluster analysis, she ends up in cluster F. People that ended up in cluster F normally are also interested in scientific books. Probability suggests that a linguistic start-up should try to sell its book to that customer. But whether this will be successful cannot be predicted on an individual basis. However, trying to sell 100 books to 100 people in cluster F should meet with some success.

The same holds true for a lexeme that ended up in a cluster N. Let the centroid of the cluster N have the value 0.93 for the dimension morphological plural. Probability is thus high that a random lexeme from that cluster will be able to mark morphological plural, and this probability is much higher than that of a random lexeme from the cluster V, where the centroid might have a value like 0.12 for the dimension morphological plural.

7 Quality of features and dimensions

An important question is the quality and validity of the features that are investigated. We can distinguish two different issues here. The first one is to determine which features should actually be analyzed (feature selection in terms of Jain et al. (1999: 271)). This is thus before the data collection. Given that vast

number of possible features, a subset has to be chosen. A first reduction can be achieved by picking features that have been proven useful in the analysis of word-class systems in the literature, or seem promising at first glance. Obvious candidates would be tense or case ,for instance. The second assessment of feature quality comes after having run the clustering algorithm (feature extraction in terms of Jain et al. (1999: 271)). There will be some features based on which a prediction of cluster membership is quite reliable, while other ones will not allow very good predictions. Features can thus provide cues to cluster membership, and the cue validity (Rosch 1981) can differ. For instance, the feature on Ø NSG in Figure 8 has a maximum cue validity, because the membership of a lexeme in the class days of the week can be predicted by looking at that feature alone. Features with a high cue validity can thus serve as a shorthand to ascertain the class-membership of a given lexeme. co-occurrence with the , for instance, should have a high cue validity for predicting (non-)membership of a given lexeme in the class of English Nouns. For reasons of convenience, a researcher can then just check the feature co-occurrence with the for a new lexeme whose class-membership is at stake and get a very good estimate.

It is important not to confound high cue validity features with definitions. Definitions are axiomatic, and cannot be proven wrong. They can just be more or less sensible. They depend on the explanatory goal that the definer wishes to achieve. Shorthands based on cue-validity on the other hand are based on data. They are independent of the explanatory goals of the analyzer.[21] Also, the cue validity of a feature can change as more features are analyzed, and so a feature that seemed very valid in the beginning can become less informative.

The "definitions" of word classes we find in grammar books are usually features with a very high cue validity. Normally, their cue validity is not 100%, which means that there are some lexemes that do not conform to all the defining features of, say, Nouns. Those are treated as nouns nevertheless, and dubbed as exceptions. These exceptions then prove that the "definitions" are in fact not definitions, but shorthands to ascertain the cluster a given lexeme belongs to.

A reviewer asks how the clustering model can account for the intuitions that native speakers have about class-membership of a given lexeme. If native speakers do indeed have intuitions about the class-membership of a given lexeme, these are probably based on features that a) have a high cue validity and are b) highly salient. An example would be Spanish verbs, which share a positive value for the feature agreement. Agreement is ubiquitous in Spanish and perceptually salient. It also

21 Other aspects of the clustering process might depend on the explanatory goal, for instance which features are selected.

separates verbs and non-verbs very neatly, resulting in a high cue validity. Intuitions about class-membership in Spanish are much more likely to be based on agreement than on other features which are less salient and have a smaller cue validity. Things are more difficult for languages like Cayuga or Samoan, where clear cues do not seem to exist.

8 Variation

We have outlined a method to compute clusters of word classes. This method consists of a distance metric, a dispersion metric, a metric for information gain/loss and a clustering method. We have taken the most simple approach to every one of these domains for illustratory purposes. These simplistic approaches invariably have their shortcomings as soon as the data are not in a class-room distribution. More sophisticated approaches exist that allow more precise clusterings or can deal with a wider range of data distributions. See Halkidi et al. (2001), Jain & Dubes (1988), Han & Kamber (2001) and Theodoridis & Koutroumbas (2006) for overviews.

9 Outlook

9.1 Fuzziness and prototypes

Research on linguistic categories has often made use of the concepts of fuzziness (Zadeh 1965) and prototypes (Rosch 1975). I will briefly discuss mathematical correlates of these notions in the mathematical model adopted here.

Fuzziness is used to describe that an item can belong to more than one category. This can easily be modeled by measuring the distance between the lexeme and the centroid of the relevant category. Suppose that the we have three clusters which present major word classes, N, ADJ and V. Let d*(stone, centroid(N)) = 0.3, d(stone, centroid(V)) = 3.4* and *d(stone, centroid(ADJ)) = 4.7*. We can interpret this as a quite strong membership of stone in the cluster N, and as a weaker membership in the clusters V and ADJ.[22]

Prototypes are defined as the most central member of a set (Rosch 1975). This is easy to apply to a cluster since we know its centroid. The prototype of a cluster can then be equated with its centroid. Since the centroid is normally not identical to any lexeme vector, the medoid can be chosen instead, which is the vector with the shortest distance to the centroid.

22 In traditional fuzzy logic, the values should add up to 1. Normalization of the distances to meet this criterion is trivial.

9.2 Relation to other theories

In this section we will see how cluster analysis compares to other theories that also make use of distributional analysis, exemplified by Dixon & Aikhenvald (2004) and Evans & Osada (2005b) and to theories with a less prominent focus on morphosyntax, namely Hengeveld's and Croft's.

Dixon & Aikhenvald (2004) claim that every language has a class of adjectives, even if it is very small: 'I suggest that a distinct word class 'adjectives' can be recognized for every human language. [...] I suggest that there are always some grammatical criteria - sometimes rather subtle - for distinguishing the adjective class from other word classes"(Dixon 2004: 1).

I agree with Dixon here, but whether the subtleness of these criteria can still be qualified by rather or whether the use of the intensifier very is more appropriate is subject to discussion. Let us discuss two examples of the subtlety of the distinction of the adjective class that Dixon offers: In Yir-Yoront (Alpher 1991), Adjectives are distinguished from Nouns by slight semantic differences that occur when they are modified by *morr* 'real, actual, very'. A second criterion is the occurrence of the postposition *mangl* 'a little', which is possible with Adjectives but not with Nouns. While these tests, if they bear out, are indeed a means of delineating two classes of words, they are very subtle, to say the least. A second language with subtle tests to distinguish the class of Adjectives, this time from the class of Verbs, is Wolof (McLaughlin 2004). Here, only the following two criteria can be used to distinguish the two classes: 1) A lexeme in intransitive predicate position within a definite relative clause untainted by a tense marker, an intensifier or a second argument is a Verb if the relative and definite markers are fused, otherwise it is an Adjective. 2) When Nouns are modified by two relative clauses, the one closer to the Noun will be the one with an Adjective in it.

Dixon takes these facts about Yir-Yoront and Wolof as support of his theory that all languages have a class of adjectives that can be defined on morphosyntactic grounds. I agree with him that the test mentioned above will single out a class in each of the two languages discussed, and that that class can very well be called "Adjectives". The question that we have to ask is: Are these classes particularly informative? Should we accept them as major categories? What is the information we gain when splitting the supercluster N/ADJ (Yir-Yoront) into two subclusters N and ADJ?[23] It is obvious that the amount of information we gain is extremely small. For the class of Adjective to be admitted as a major word class in the two languages under discussion, we must choose a value for our threshold Θ or τ that is very low. It is extremely likely that such a low value for τ will not only single out a separate

23 V/ADJ in V and ADJ in Wolof

class of Adjectives as a major lexical category. It would also separate count nouns from mass nouns, stative verbs from active verbs, chronometrical expressions of time from diurnal ones (Figure 8) and so on,[24] all of those as major lexical categories on a par with Adjectives. This seems not be desirable. As a working principle, the value for τ should not be chosen in order to single out one's favorite class but should either be predetermined at a fixed value for cross-linguistic comparison, or be determined by the 'knee' method when considering only one language. We can conclude that while Yir-Yoront and Wolof certainly have a class of Adjectives that can be defined by morphosyntactic criteria, this class does not have the special status that the major lexical categories enjoy. It is rather a very low subclass of a major lexical category.[25] While Dixon suggests that criteria for singling out adjectives can be found in any language, he does not imply that these adjectives all have to behave in an absolutely identical way. He refers to Corbett (2004), who shows that Russian adjectives can be defined by five criteria, but actually only very few of them meet all five criteria. We can see this as an application of the principle outlined above that elements in a cluster need not be completely identical in their distribution as long as they are more similar among themselves than they are to elements outside the cluster.

A different stance on this is taken by yet other advocates of the distributional method, Evans & Osada (2005b). They advocate three principles for establishment of morphosyntactic categories: *distributional equivalence, semantic compositionality* and *bidirectionality*. Distributional equivalence means that all lexemes in one class must have the absolute same behaviour, semantic compositionality means that the effects of conversion have to be predictable and bidirectionality means that conversion must not be a unidirectional process. In relation to the methodology exposed in this paper, the last criterion is something that is irrelevant for cluster analysis, the second criterion cannot be applied to cluster analysis as long as there is no formalism to compute identity of semantic behaviour under conversion[26] and the first criterion is extremely problematic.

In clustering terms, absolute distributional equivalence means that the distance between two vectors in one cluster may be no greater than 0. Clustering is only allowed if the distance is 0. This boils down to saying that one should not

24 Under the assumption that there will be at least two criteria that separate all of the pairings cited. I am confident that at least two criteria can be found for the cited examples.

25 From what I understand, Dixon is actually not opposed to seeing adjectives as a subclass of a bigger class given his discussion of the Lao situation (Enfield, 2004), where we find Adjectives as a subclass of Stative Verbs, themselves a subclass of Verbs.

26 And Evans does not say how this identity can be established or give a reference to where such a procedure is explained.

cluster. Even the slightest deviation would discard the possibility that two lexemes end up in the same cluster. Coming back to Crystal's data, that would mean that for the 26 temporal nouns discussed above more than 20 separate classes have to be assumed because they are not distributionally equivalent. In light of Gross's study, it is even more utopian to demand distributional equivalence.[27] If one is serious about absolute distributional equivalence for every member of a lexical category, one would have to assume 12000 major lexical categories alone for the lexemes formerly known as French Verbs. We conclude that the criterion of absolute distributional equivalence cannot be upheld as such. However, relative distributional equivalence seems a reasonable thing to demand: All lexemes in one class should be more similar among themselves than they are to lexemes outside their class. Cluster analysis is able to model this. We have seen that the threshold Θ measures the absolute amount of dispersion that is permitted within a cluster. When set to 0, no dispersion is allowed, which means very small classes. Higher values then yield bigger and bigger classes. More research is needed to find a sensible value for Θ , but I propose a value of normalized Θ=0.4 to start with. That means that the total dispersion found in the data set under discussion is set to 1, and no single cluster may have an internal dispersion greater than 40% of the total dispersion of the data.

Absolute distributional equivalence, as required by Evans' first criterion, cannot be upheld. Evans' second criterion is semantic rather than morphosyntactic, but suffers from the same requirement of absolute identity. It states "there should be isomorphic semantic changes in all lexemes placed in a given functional position"[370]. Just as the morphosyntactic behaviour is never exactly the same, it is likely that also semantic effects of conversion might show very subtle differences between any two lexemes. The three verbs of ingestion *drink*, *eat*, *smoke* are arguably very similar in their semantics. Yet conversion does not apply in the same way to all of them. Compare *They drink/eat/smoke, They have a drink/* an eat/a smoke, There is a drink/*an eat/*a smoke on the table.* If Evans is serious about the criterion of identity of semantic change under conversion, drink, eat and smoke have to be considered as belonging to different word classes. If one does not want to throw out the baby with the bath water, it will be wise to allow for semantic differences up to a certain threshold along the criteria outlined above for morphosyntactic distribution.[28]

The theories discussed above are based on distributional analysis in one

27 This is already noted in Croft's reply to Evans in the same issue.

28 A mathematical measure of semantic distance seems more remote that a measure of morphosyntactic difference, but this does not change the necessity of such a threshold if this second criterion is to be upheld.

particular language at a time in order to arrive at a good description of the lexical categories of that language. These approaches are data-driven, language-specific and rely mainly on morphosyntax. While having the advantage of not coercing an individual language into categories that might not suit it, this language-particular procedure makes cross-linguistic generalizations very hard, a fact that Evans & Osada (2005a) explicitly admit. Before analysing theories that go beyond pure morphosyntactic distribution, let us see how cluster analysis can be applied to cross-linguistic generalizations.

Like the other theories, cluster analysis cannot help in establishing the identity of word classes in different languages. It is subject to the same limitations as the classical distributional analysis. Where cluster analysis has an edge over the classical analysis is in determining the number of word classes. It is possible to state that language A has a 2-class system, a 3-class system, or even a 2(4)-class system for given values of Θ and τ , compare Figure 7. Cluster analysis thus gives information about the number of cluster, but not about their semantics.[29]

It is then possible to a) compare the number of word classes obtained for different languages and b) relate the number of word classes to other typological parameters, for instance word order. Cluster analysis is thus an approach that makes use of language particular constructions, but whose results can be compared cross-linguistically, bridging the gap between language description and typology.

Other frameworks that do not limit themselves to one particular language but strive for cross-linguistic comparability are spearheaded by Hengeveld and Croft. These theories rely on facts other than morphosyntactic analysis. What is the relationship between cluster analysis and these theories?

Hengeveld[30] proposes a four-fold division of parts-of-speech based on their occurrence in predicate phrases or term phrases, and their functions as head or modifier. Parts-of-speech are thus not arrived at inductively by their morphosyntactic distribution, but are defined by the propositional functions they fulfill.[31] It would be interesting to square the clustering approach and the functional approach for the same language. The prediction would be that none of the emerging major word classes would cross-cut classes defined by Hengeveld. E.g. we would not expect to find one lexeme that can only be the head of a predicate ($V_{Hengeveld}$) in

29 This does not exclude that a linguist might investigate a cluster and find semantic regularities in it. But this is a step that follows up on cluster analysis and is not included in it.

30 Hengeveld (1992a), Hengeveld et al. (2004), Hengeveld (1992b), this volume

31 A precise propositional function is possibly determined by morphosyntactic criteria. In our discussion of the theories, we only compare the starting principles that the theories are based on, and not other methods they draw on at a certain point. Hengeveld's starting point is the propositional function.

our cluster A and another $V_{Hengeveld}$ in a cluster B. Co-occurrence of a Hengeveldian class with other Hengeveldian word classes in one cluster is no problem, though. Since Hengeveld's word classes are not predictions but definitions, they cannot be falsified. But they can be evaluated on the base of their sensibility. If the definitions appear to match reality, they are sensible, if they are in conflict with reality, other definitions should be looked for. Cluster analysis can then be used as a tool to test whether Hengeveld's definitions match reality.

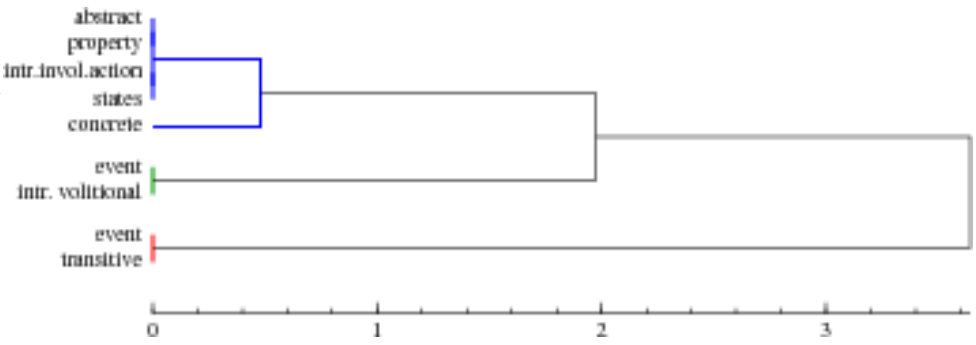


Figure 11: Clustering of Guarani lexemes, schematic

Croft[32] claims that word classes cluster around universal prototypes and that every language's parts-of-speech system is an instantiation of this universal prototype. The predicted prototypes are N for referring objects, ADJ for modifying properties and V for predicating actions. If this theory is correct, we expect the major word classes established by clustering to always be three or less. Systems with four or more clusters should not exist, discarding the types a), b) and possibly c) in Figure 7. Furthermore, every cluster should instantiate the universal prototypes. This means, just as with Hengeveld's theory, that no language should have some N_{Croft} in one cluster and some N_{Croft} in another one. Having, say, N_{Croft} and V_{Croft} in one cluster is no problem.

While Croft's prediction is certainly true for most of the world's languages, the clustering method provides a counter-example for the Guarani parts-of-speech system (Nordhoff 2004). A reduced dendrogram is shown in Figure 11. There are three clusters, but these do not seem to instantiate Croft's universal prototypes. A first separation is made on the grounds of transitivity and a second one based on volitionality. This yields a class of transitive lexemes, a class of intransitive but volitional lexemes and a class for intransitive involitional lexemes. In every class, we find some V_{Croft} (*juka* 'kill' in the transitive class, *guata* 'walk' in the volitional

32 Croft (2001), Croft (2000), Croft (1991)

class, *atɲa* 'sneeze' in the involitional class).[33]

10 Conclusion

It was also one of Croft's claims that gave rise to this paper: the one that there is no way to decide when to stop splitting. I hope to have shown that this claim does not hold. There are two ways two determine when to stop: either set Θ and τ to a fixed value before the clustering process starts, or use the 'knee' method as a more flexible means. This means that at least the number of word classes can cross-linguistically be established on hard and fast grounds. This number can then be related to other typological parameters, yielding testable hypotheses like e.g. *1-class languages (determined with Θ=0.4) prefer prefixation over suffixation* or *3-class languages (determined with Θ=0.3) are more often ergative than chance would predict* etc.

While the quantity of clusters can be used as a parameter for typological generalizations, the semantic cross-linguistic comparability of word classes does not improve through cluster analysis. Generalizations like *Languages with a class of adjectives tend to have property X* will not be any more empirically grounded than they were before.

Bibliography

Aarts, Bas. 2006. “Conceptions of categorization in the history of linguistics”. *Language Sciences* 28 :361–385.
Alpher, Barry. 1991. Y*ir-Yoront lexicon: sketch and dictionary of an Australian language.* Berlin: Mouton de Gruyter.
Bhat, DNS. 1994. *The Adjectival Category: Criteria for Differentiation and Identification.* Amsterdam: Benjamins.
Boberg, Jorma & Tapio Salikowski. 1993. “General formulation and evaluation of agglomerative clustering methods with metric and non-metric distances”. *Pattern Recognition* 26 (9):1395–1406.
Broschart, Jürgen 1997. “Why Tongan does it diffferently”. *Linguistic Typology* 1 (2):123–166.
Clark, E. & H. Clark. 1979. “When nouns surface as verbs”. *Language* 55 :767–811.
Comrie, Bernard & Petra Vogel (eds.). 2000. *Approaches to the Typology of Word Classe*s. Berlin and New York: Mouton de Gruyter.
Corbett, Greville. 2004. “The Russian Adjective: A pervasive yet elusive category”. In Dixon & Aikhenvald. (2004), 199–222.
Croft, William. 1991. *Syntactic Categories and Grammatical Relations: The Cognitive Organization of Information*. Chicago: University of Chicago Press.

33 Note that Hengeveld's approach can deal with this if his definitions are applied rigorously. In that case, only the transitive class are $V_{Hengeveld}$, while the other two classes are found together in $Non\text{-}Verb_{Hengeveld}$. It might sound strange to call *guata* 'walk' a 'Non-Verb', but nothing in Hengeveld's theory precludes this.

Croft, W. 2000. “Parts of speech as language universals and as language-particular categories”. In Comrie & Vogel (2000), 47–64.
Croft, William. 2001. *Radical Construction Grammar* . Oxford: OUP.
Croft, William. 2005. “Word classes, parts of speech, and syntactic argumentation”. *Linguistic typology* 9 (3):431–441.
Crystal, David. 1967. “Word Classes in English”. *Lingua* 67 :24–56.
Dixon, RMW. 2004. “Adjectival classes in typological perspective”. In Dixon & Aikhenvald (2004), 1–49.
Dixon, RMW. & Alexandra Aikhenvald (eds.). 2004. *Adjective classes : a cross-linguistic typology.* Oxford: OUP.
Eisen, Michael. B., Paul. T. Spellman, Patrick. O. Brown & David. Botstein. 1998. “Cluster analysis and display of genome-wide expression patterns”. *Proc. Natl. Acad. Sci.* USA 95 :14 863–14 868.
Enfield, Nick. 2004. “Adjectives in Lao”. In Dixon & Aikhenvald (2004), 323–347.
Evans, Nick. & Toshiki Osada. 2005a. “Mundari and argumentation in word-class analysis”. *Linguistic typology* 9 (3):442–457.
Evans, N. & Toshiki Osada. 2005b. “Mundari: The myth of a language without word classes”. *Linguistic typology* 9 (3):351–390.
Givón, Talmy. 1984. Syntax: *A Functional-Typological Introduction*. Amsterdam: Benjamins.
Gross, Maurice. 1979. “On the failure of generative grammar”. Language 55 :859–885.
Guha, S., R. Rastogi & K. Shim 1998. “CURE: An Efficient Clustering Algorithm for Large Databases”. In *Proceedings of the ACM SIGMOD Conference.*
Guha, Sudipto, Rajeev Rastogi & Kyuseok Shim. 1999. “ROCK: A Robust Clustering Algorithm for Categorial Attributes”. In *Proceedings of IEEE Conference on Data Engineering.*
Guo, Diansheng, Donna J. Peuquet & Mark Gahegan. 2003. “ICEAGE: Interactive Clustering and Exploration of Large and High-Dimensional Geodata”. *Geoinformatica* 7 (3):229–253.
Halkidi, Maria, Yannis Batistakis & Michalis Vazirgiannis. 2001. “On clustering validation techniques”. *Journal of Intelligent Information Systems 17* (2/3):107–145.
Han, Jiawei & Micheline Kamber. 2001. *Data Mining – Concepts and Techniques.* San Francisco:Morgan Kaufmann.
Hengeveld, Kees. 1992a. “Non-verbal predicability”. In Kefer, Michel & Johan van der Auwera (eds.). *Meaning and Grammar* , Berlin, New York: Mouton de Gruyter, 77–94.
Hengeveld, Kees.1992b. “Parts of speech”. In Fortescue, Michael; Peter Harder & Lars Kristoffersen (eds.). *Layered Structure and Reference in a Functional Perspective*, Amsterdam: Benjamins. 29–56.
Hengeveld, Kees & Jan. Rijkhoff. 2005. “Mundari as a flexible language”. *Linguistic typology* 9 (3):406–431.
Hengeveld, Kees, Jan Rijkhoff & Anna Siewierska .2004. “Part-of-Speech Systems and Word Order”. *Journal of Linguistics* 40 (3):527–570.
Hopper, Paul & Sandra Thompson. 1984. “The Discourse Basis for Lexical Categories and Universal Grammar”. *Language* 60 :703–752.
Jain, A. K., M. Murty & P. Flynn. 1999. “Data Clustering: A Review” *ACM Computing Surveys,* 31 (3):264–323.
Jain, A. K. & R. C. Dubes. 1988. *Algorithms for clustering data*. Englewood Cliffs: Prentice Hall. McLaughlin, Fiona. 2004. “Is there an Adjective Class in Wolof?” In Dixon & Aikhenvald (2004), 242–262.
Mithun, Marianne. 2000. “Noun and verb in Iroquoian languages:”. In Comrie & Vogel (2000),

379–420.
Mosel, Ulrike. & Even Hovdhaugen. 1992. *Samoan Reference Grammar. Oslo*: Scandinavian University Press.
Nordhoff, Sebastian. 2004. *Nomen/Verb-Distinktion im Guarani* . Arbeitspapier Neue Folge Nr. 48, Köln: Universität zu Köln.
Pal, Nikhil R. and Sankar K. Pal. 1993. "A review on image segmentation techniques". *Pattern Recognition* 26 (9):1277– 1294.
Quirk, Randolph. 1965. "Descriptive Statement and Serial Relationship". *Language* 41 (2):205–217.
Ray, Siddeshwar. & Rose H. Turi. 1999. "Determination of Number of Clusters in K-Means Clusteringand Application in Colour Image Segmentation". In *Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques*. 137–143.
Rosch, Eleanor. 1975. "Cognitive representation of semantic categories". *Journal of Experimental Psychology*: General 104 :192–233.
Rosch, Eleanor. 1981. "Human Categorization". In Warren, Neil (ed.). *Studies in Cross-cultural psychology*, London: Academic Press. 1–49.
Ross, John R. 1972. "The Category Squish: Endstation Hauptwort". In Peranteau, Paul M; Judith N. Levi & Gloria C. Phares (eds.). *Papers from the Eighth Regional Meeting. 2*, Chicago: Chicago Linguistic Society, 316–339.
Salvador, Stan. & Philip Chan. 2004. "Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms". In 1*6th IEEE International Conference on Tools with Artificial Intelligence.* 576– 584.
Sasse, Hans-Jürgen. 1988. *Der irokesische Sprachtyp*. Köln: Institut für Sprachwissenschaft.
Sasse, Hans-Jürgen. "Das Nomen – eine universale Kategorie?" *Sprachtypologie und Universalienforschung* 46 :187–221.
Schone, Patrick & Daniel Jurafsky. 2001. "Language-independent Induction of Part of Speech Class Labels Using Only Language Universals". Ms.
Seuren, Pieter. A. M. 1998. *Western Linguistics – An historical introduction*. Oxford: Basil Blackwell.
Sharma, Subhash C. 1996. *Applied multivariate techniques*. London: John Wiley and Sons.
Theodoridis, Sergios & Konstantinos Koutroumbas. 2006. *Pattern Recognition*. Academic Press, 3rd edition.
Ushioda, Akira. 1996. "Hierarchical Clustering of Words." In *16th International Conference on Computational Linguistics, Proceedings of the Conference.* 1159–1162.
Walsh, Michale. 1996. "Vouns & Nerbs: A Category Squish in Murrinh-Patha". In McGregor, William (ed.). *Studies in Kimberley Languages in honor of Howard Coate*, München: LINCOM. 227–252.
Wang, Wen & Dimitra Vergyri. 2006. "The Use of Word N-grams and Parts of Speech for Hierarchical Cluster Language Modeling." *In Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing.*
Wetzer, Harrie. 1996. *The Typology of Adjectival Predication*. Amsterdam: Benjamins.
Zadeh, Lotfi. A. 1965. "Fuzzy sets". *Information and control* 8 :338–353.