CADERNOS
de_LINGUÍSTICA

cadernos.abralin.org

RESEARCH REPORT

QUANTIFYING THE DIFFERENCES BETWEEN LEXICAL CATEGORIES: THE CASE OF PRONOUNS AND DETERMINATIVES IN ENGLISH

Brett REYNOLDS
Humber College

ABSTRACT

The Cambridge grammar of the English language (HUDDLESTON; PULLUM, 2002) attempts to present a comprehensive and rigorous description of Modern Standard English. Much of the book is taken up with describing the properties of the various lexical categories, including determinative and pronoun. The distinction between these categories has been questioned by various authors in English (ABNEY, 1987; CROFT, 2001; HUDSON, 2004; MATTHEWS, 2014; POSTAL, 2014/1966; SOMMERSTEIN, 1972) and other languages (e.g., NAU, 2016). Here, I employ energy distance, a novel family of non-parametric statistics, to adjudicate between these positions. Following Crystal (1967), I binarily encode the features (has/doesn't have feature) of the determinatives and pronouns from CGEL in a 138 word-forms by 232 features matrix. The results provide support for CGEL's analysis (*k*-groups produces a 93% correspondence with CGEL's categorization) and show that energy distance statistics applied to such matrices can help us adjudicate between competing lexical category analyses without resorting to methodological opportunism (CROFT, 2001).

OPEN ACCESS

EDITORS
– Miguel Oliveira, Jr. (UFAL)
– René Almeida (UFS)

REVIEWERS
– Guilherme Fromm (UFU)
– João Cyrino (UFBA)

DATES
– Received: 30/07/2021
– Accepted: 17/08/2021
–Published: 08/27/2021

HOW TO CITE
REYNOLDS, Brett (2021). Quantifying the Differences Between Lexical Categories: The Case of Pronouns and Determinatives in English. *Cadernos de Linguística*, v. 2, n. 3, e399.

RESUMO

The Cambridge grammar of the English language (HUDDLESTON; PULLUM, 2002) tenta apresentar uma descrição abrangente e rigorosa do inglês padrão moderno. Muito do livro é dedicado à descrição das propriedades das várias categorias lexicais, incluindo determinativos e pronomes. A distinção entre essas categorias foi questionada por vários autores em inglês (ABNEY, 1987; CROFT, 2001; HUDSON, 2004; MATTHEWS, 2014; POSTAL, 2014/1966; SOMMERSTEIN, 1972) e outras línguas (por exemplo, NAU, 2016). Aqui, eu emprego energy distance, uma nova família de estatísticas não paramétricas, para julgar entre essas posições. Seguindo Crystal (1967), codifico binariamente as características (tem / não tem característica) dos determinativos e pronomes do CGEL em uma forma de 138 palavras por 232 características matriciais. Os resultados fornecem um suporte para a análise do CGEL (*k*-groups produzem uma correspondência de 93% com a categorização do CGEL) e mostram que as estatísticas de energy distance aplicadas a tais matrizes podem nos ajudar a decidir entre análises de categorias lexicais concorrentes sem recorrer ao oportunismo metodológico (CROFT, 2001).

KEYWORDS

Lexical Categories; Parts of Speech; Determinatives; Pronouns; Energy Distance; Methodological Opportunism.

PALAVRAS-CHAVE

Categorias Lexicais; Partes do Discurso; Determinativos; Pronomes; Energy Distance; Oportunismo Metodológico.

INTRODUCTION

Lexical categories are a matter of perennial dispute in linguistics. This is true as much for descriptive categories as for comparative concepts (HASPELMATH, 2018) and applies regardless of how carefully studied the language is. In 1967, Crystal concluded that "word classes in English are more complex things than is still generally supposed; and … before we can produce a set of satisfactory definitions, we need to examine the distribution of single words much more thoroughly" (CRYSTAL, 1967, p. 55). Many would say that despite the intervention of over half a century, lexical categories are still more complex things than is generally supposed and that we still lack a set of satisfactory definitions, even for the lexical categories of English, one of the most comprehensively studied languages.

Of the lexical categories, perhaps determinatives and pronouns throw up the most disagreement. "The connection between pronouns and determinatives is striking in many languages. If they are not to be grouped into one class, their relationship must be clarified"[1] (NAU, 2016, p. 96). And this is true even English.

> Although the term 'determiner' [or the term I use here, determinative] has been in common use for nearly half a century, there is no consensus in detail, especially among recent authorities, either as to what words should be assigned to such a part of speech or, if categories and functions are distinguished, what the units are that have a 'determining' function. This is true both intensionally, of the features by which determiners have been defined, and extensionally, of the range of individual words included. (MATTHEWS, 2014, p. 69)

Here the greatest confusion is between determinatives and pronouns: is *my* a determinative because it appears before a noun, helping us to pick out its referent, or is it a pronoun because of the morphological similarities it shares with other pronouns? Is *many* always a determinative or is it a pronoun when it stands alone as a subject (e.g., *many were swayed*)? The Cambridge grammar of the English language (HUDDLESTON; PULLUM, 2002, hereafter CGEL), attempts a detailed and systematic answer to these questions. It describes the determiner function, and the categories of determinative and pronoun in exquisite detail over almost 100 pages in chapter 5, starting on page 354. Yet doubt and disagreement remain.

Working in early transformational theories, POSTAL (2014/1966) argues "the so-called pronouns *I*, *our*, *they*, etc. are really articles, in fact types of definite article. However, article elements are only introduced as segments in intermediate syntactic structures" (p. 15). SOMMERSTEIN (1972) makes the converse claim, but both deny a fundamental difference between determinatives and pronouns. ABNEY (1987) takes up Postal's arguments and

[1] "Der Zusammenhang zwischen Pronomina und Determinativen ist in vielen Sprachen auffällig. Wenn sie nicht zu einer Klasse zusammengefasst werden sollen, muss doch ihre Beziehung geklärt werden." (my translation)

develops them in his so-called "DP analysis". HUDSON (2004), in a more modern theory, word grammar (HUDSON, 2007), dispenses with the deep-structure arguments and straightforwardly argues that determinatives "are a subset of pronouns ranging over many of the traditional pronoun types — demonstrative, possessive, interrogative and so on" (p. 10). CROFT (2001) goes a step further and actually calls the whole enterprise into question, accusing researchers attempting to establish lexical categories of "methodological opportunism," and claiming, "there is no a priori way to decide which of several constructions with mismatching distributions, or which subset of constructions, should be chosen as criteria for identifying the category in question" (p. 41).

CROFT's observation highlights a difficulty with the way the enterprise of lexical categorization has been conducted, which is through logical argumentation. Such an approach could be definitive if necessary and sufficient conditions were identified, but this has not been possible; exceptions abound, and consensus is far from forthcoming. So, as CRYSTAL says, "the only realistic solution seems to be statistical" (1967, p. 45). This would involve not picking and choosing from among the possible criteria, but rather setting out all the likely[2] criteria and then analysing them.

The first step in such an approach was described by CRYSTAL (1967). An example of what he had in mind can be seen in Table 1, reproduced in part below, where he set out a matrix with rows for "adverb" lexemes and columns for the following criteria:

(1) ability to occur immediately before or after verb, viz. Subject (Adverb) Verb (Adverb)

(2) ability to take intensifier without preceding determiner

(3) ability to occur initially (mobility criterion) in sentence.

	1	2	3
asleep	+	+	+
inside, downstairs	–	–	+
alike	+	+	? –

Table 1. A matrix of words and features with binary categorical coding (reproduced from CRYSTAL, 1967, p. 52)

The criteria here are syntactic, but there is no a priori reason to rule out phonological, grammatical, lexical, or semantic criteria (CRYSTAL, 1967).

What CRYSTAL had in mind, though, is something like a simple tally. "One would always expect a coherent word class to have at least one criterion with 100% applicability, to justify one's intuition of coherence" (p. 45), and then other criteria could be ranked by the number of words to which they apply. I reject that approach as unworkable. There are almost never

[2] Of course, there is no end to the possible criteria. We could consider all words appearing paragraph initially in the first British printing of *Moby Dick*, but the idea is to be inclusive without becoming absurd.

criteria with 100% applicability when it comes to the lexical categories that linguists, lexicographers, language teachers, and language learners find most useful. But lexical categories are useful to the extent that they help us predict features of their members, even if they are not perfectly reliable. More importantly, the more features that a given lexical category suggests, the more useful that categorization is.

Statistics can help us to evaluate our categorizations by comparing the probability of a set of features given a certain categorization to their probability under a random distribution. This is usually expressed as a *p*-value, ranging from 1 (indistinguishable from random) to 0 (impossible under a random distribution). We can then argue that, allowing for some error, a categorization that leads to a lower *p*-value over a given set of features is better than one that leads to a higher value.

Another approach is to use an unsupervised learning algorithm to analyze the data and create clusters of words. These clusters could then be compared against a hypothesis. Clustering does not produce a *p*-value, but it can produce plots that can be examined to see if the groupings are like the categories and sub-categories set out in a grammar such as CGEL.

1. ENERGY DISTANCE STATISTICS

Energy distance is a recently developed family of non-parametric statistics (RIZZO; SZÉKELY, 2016). There is also a related package for R (R CORE TEAM, 2019) called energy (RIZZO; SZÉKELY, 2019), which allows researchers to calculate these statistics relatively easily.

Statistics can be classified as parametric or non-parametric. In parametric statistics, a model is selected to reflect some assumptions about your data. An example might be a linear model. Once a model has been selected, the degree of fit between the model and the data is calculated. This is done because “it is generally much easier to estimate a set of parameters … than it is to fit an entirely arbitrary function” (JAMES, 2017, p. 22). Parametric statistics require relatively few observations compared to non-parametric statistics. Because of these advantages, many parametric options are available, and parametric statistics are often preferred by researchers.

However, in many cases the assumptions cannot be met. Models commonly assume, for example, that the data is normally distributed or that there is constant variance. In the case of matrices like that in Table 1, the data is in the form of binary categorical variables (e.g., has genitive case +/−), instead of discrete or continuous variables, so the error is neither normally distributed, nor is there constant variance. A case like this requires non-parametric statistics, which “seek an estimate of *f* that gets as close to the data points as possible without being too rough or wiggly" (JAMES, 2017, p. 23). Unfortunately, fewer non-parametric options exist, and they are often less familiar to researchers, but energy distance is now available.

The analysis I would like to conduct here includes multiple features (e.g., has genitive case, starts with /ð/, marks an NP as definite, etc.), which is to say that it is a multivariate analysis. Typically, MANOVA would be applicable for this kind of situation, but MANOVA won't work because it is a parametric model requiring "that random error is normally distributed with mean zero and constant variance" (RIZZO; SZÉKELY, 2010, p. 1034). This is precisely the kind of situation where energy statistics will work. Also, the common problem of too few observations underpowering non-parametric statistics is not an issue because the list of words can include the entire universe of CGEL determinatives and prepositions.

As far as I can discern, energy statistics have not been previously employed in linguistics research, perhaps because they have not been available for long. This study will then be an evaluation of these statistics for lexical categorization as well as an evaluation of the categories set out in CGEL. This weighing of two approaches to categorization is akin to the epistemology behind statistics like inter-rater reliability estimates. In both cases, we cannot be sure of the validity of any individual rater's judgement, but if multiple raters converge on similar judgements, then this has been seen as the basis for taking their judgements to be reliable, reliability being one aspect of construct validity (MESSICK, 1995). Similarly, if CGEL categories are well supported by energy statistics, this provides evidence to support the validity of each.

1.1. K-GROUPS CLUSTERING

The *k*-groups unsupervised learning algorithm developed by LI (2015) is part of the energy distance family. It is a generalization of the more familiar parametric *k*-means algorithm. *K*-means measures the statistical distance between pairs of clusters and searches for "the best partition which maximizes the total between-clusters energy distance" (LI, 2015). Typically, *k*-means would be applicable for discovering clusters like lexical categories, but, like MANOVA, *k*-means will not work with the data used here for the reasons explained above (LI, 2015, p. 49). The *k*-groups method overcomes these limitations.

Once the words have been assigned to clusters by *k*-groups, it is possible to compare those clusters to CGEL's categories to see how much overlap exists.

1.2. VISUALIZATION AND DENDROGRAMS

The energy algorithm for clustering is formally similar to Ward's method (RIZZO; SZÉKELY, 2016). Where the *k*-groups algorithm is designed to maximize the distance between the two groups, the algorithm for clustering starts with observations of single pairs of words, and, at each step, merges clusters that have minimum cluster distance (RIZZO; SZÉKELY, 2016, p. 30). This results in hierarchically structured sets of pair-wise clusters. The resulting structure can be plotted as a dendrogram, which provides a way to see detailed similarities between words. Because of the way this algorithm proceeds, the largest two groups may not be maximally distant in the way that they would be under *k*-groups. The purpose, therefore, is to observe the detailed structure to see which words and clusters of words are closest together.

1.3 DISTANCE COMPONENTS (DISCO)

Distance components (DISCO; RIZZO; SZÉKELY, 2010) is a member of the energy family of non-parametric statistics. Unlike the first two procedures, it doesn't discover the structure in the data but rather takes whatever two groups it is given and calculates both the between-groups distance and the within-group distances. These two measures can then be used to calculate an "F" ratio, which, according to EVERITT (2006), is calculated as

$$F = \frac{\text{between-groups variability}}{\text{within-group variability}}$$

A larger *F* indicates that the two categories are more distinct, relative to their internal differences. Unfortunately, there is no commonly known average or minimum *F* value when it comes to lexical categories in English or in any language. As a result, an *F*-ratio calculated on the CGEL pronouns and determinatives is currently uninterpretable on its own. Nevertheless, a *p*-value can be derived from *F*. Conventionally, any *p*-value smaller than 0.05 is considered to show that the categorization is significantly different from random.

Even a statistically significant *p*-value, though, will not be enough to confirm that two groups of words should be considered distinct lexical categories. It may be the case that even subcategories may produce significantly large *F*-values.[3]

2. METHOD

2.1. RESEARCH QUESTIONS

I have five research questions:

[3] I'd like to thank reviewer Prof. Dr. João Paulo Cyrino for this observation and for some helpful suggestions about R coding.

(1) How well do CGEL's pronouns and determinative categories align with the clusters derived from *k*-groups?

(2) How well does the hierarchical structure of the dendrogram match our intuitive notions of the similarities within the pronouns and determinative categories?

(3) What is the *F*-ratio of CGEL's pronouns and determinatives resulting from the DISCO test?

(4) Is the *F*-ratio statistically significant?

(5) If the *F*-ratio from (4) is statistically significant, are *F*-ratios calculated on subsets of CGEL's pronouns and determinative categories also statistically significant?

2.2. PREPARING THE DATA MATRIX

To begin answering these questions, I constructed a matrix like the one in Table 1 (downloadable as REYNOLDS, 2021). It ended up with 138 rows for word forms (73 determinatives and 65 pronouns) and 232 columns for features, and the aim was to be as inclusive as possible. The list of word forms is based on the CGEL descriptions of determinatives and pronouns, and it includes all words specifically mentioned as pronouns or determinative in CGEL. I made the choice to use word forms, rather than lexemes because, as NAU (2016) says, "it is not lexemes that are used in syntactic functions, but word forms"[4] (p. 31).

Next, I built the list of features. These were grouped as morphological (139), phonological (3), semantic (36), and syntactic (54) features. In the morphological group, there was a column for each word appearing as part of another word (e.g., *any* appears in *any*, *anybody*, *anyone*, *anything*, and *anywhere*). Almost all the other features were taken from CGEL, but other sources were included where a particular concept seemed relevant. Sometimes these came from the literature (e.g., must be outranked by a coindexed element, SAG; WASOW; BENDER, 2003, p. 292) and sometimes they were just features that struck me as possibly relevant (e.g., starts with /ð/).

Finally, I coded each cell in the table as "may exhibit the feature" or "never does". In many cases, I relied on my own judgement, often informed by corpus queries. With over 30,000 cells to deal with, it was impractical to do anything else. I have made my data available, and any researcher who finds errors or disagreements is welcome to publish their revisions.

[4] "Denn nicht Lexeme werden in syntaktischen Funktionen gebraucht, sondern Wortformen." (my translation)

2.3. DATA ANALYSIS AND HYPOTHESES

The first step in the analysis was to cluster the data with *k*-groups and to compare the resulting clusters with the CGEL categories. I expected to find a high degree of overlap. Second, I created a dendrogram and visually inspected it to understand the hierarchical structure of the data. I expected to find intuitive structure. Third, I ran the DISCO analysis on the full set of data grouped into CGEL determinatives and pronouns. I expected to find a significant difference between the two categories. Fourth, I ran the DISCO analysis on the determinatives grouped into the first 36 determinatives listed alphabetically and the remaining 37. I expected to find a smaller, non-significant *F*-ratio. I used the Energy package (RIZZO; SZÉKELY, 2019) in R (R CORE TEAM, 2019) to perform for all analyses.[5]

3. RESULTS

3.1. K-GROUPS CLUSTERING

The clusters resulting from the *k*-groups were very similar to the CGEL categories. Only three CGEL determinatives were assigned to the *k*-groups pronoun cluster, the personal determinatives *you*, *we*, and *us* (as in *<u>you</u> kids can come too*). And only six CGEL pronouns were assigned to the *k*-groups determinatives cluster. These include one reciprocal pronoun *one another* (though not *each other*), the two dummy pronouns *it* and *there*, and the interrogative and relative pronouns *what* and *whatever*, along with relative *which*. Overall, then, there was agreement on 129 out of 138 words (93.48%). A random assignment would centre on 50%, so, to correct for this, 50% can be subtracted and then the result doubled, which comes to a 86.96% “adjusted agreement” value. So, the answer to research question 1 is “very well”, which is in line with my expectations.

3.2. VISUALIZATION WITH DENDROGRAMS

The dendrogram is reproduced as Figure 1. The CGEL pronouns are enclosed in the upper red rectangle, and the CGEL determinatives are enclosed in the lower blue triangle. The two dummy pronouns *it* and *there* are enclosed in a smaller red triangle inside the blue triangle.[6]

[5] See the appendix for the R code.
[6] These rectangles were added manually after the dendrogram was produced.

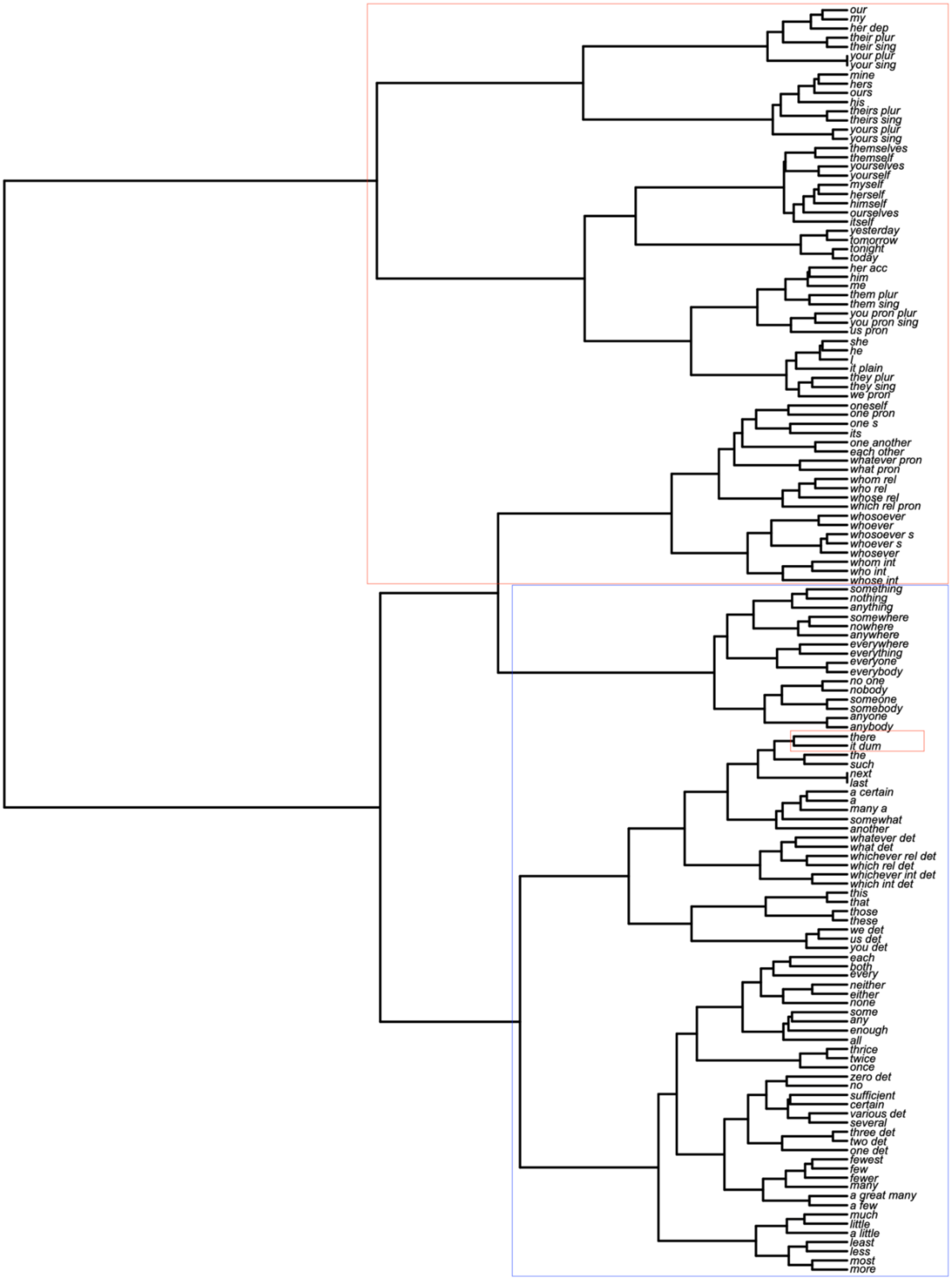

Figure 1. Dendrogram showing structure of the pronoun and determinative categories.

There is a good deal that is very intuitive about the hierarchical structure here. To demonstrate, I will describe the pronoun structure using CGEL terminology. All the genitive pronouns live together on one branch at the top of the dendrogram. This splits into a branch with seven dependent genitive pronouns (e.g., *my time*) and a branch with eight independent genitive pronouns (e.g., *mine*), the odd one out being *his*, which is both dependent and independent. The next group is the reflexive pronouns (e.g., *myself*). Somewhat unintuitively, these share a branch with the temporal pronouns *today*, *tomorrow*, *tonight*, and *yesterday*. These are followed by a branch that splits into the accusative pronouns and the nominative pronouns. This completes the top branch of the dendrogram, but 22 CGEL pronouns are clustered on branches with determinatives. The majority of these are interrogative and/or relative pronouns, which all share a branch along with the pronouns *oneself*, *one*, *one's*, and *its*, along with the reciprocals *one another* and *each other*. Inside the determinatives group, we find similar levels of structure with groupings of interrogatives, demonstratives, quantifiers, and others. So, the answer to research question 2 is also "very well", which is in line with expectations.

3.3. DISCO

The results of the DISCO test conducted with the full set of words and features are presented in Table 2.

Source	Df	Sum Dist	Mean Dist	*F*-ratio	*p*-value
factors	1	30.58286	30.58286	11.670	0.001
Within	136	356.41607	2.62071		
Total	137	386.99893			

Table 2. DISCO test output.

The mean distance between the determinatives and the pronouns is 30.58,[7] while the mean distance within each group is only 2.62. The *F*-ratio is simply the ratio of these two distances, as described in §2.3, so the mean distance between categories is 11.67 times more than the mean distance within the categories. This *F*-ratio is significant at $p \leq 0.001$, strongly suggesting that CGEL's determinatives and pronouns are indeed significantly different groups of words. This is in line with expectations.

When I conducted the DISCO test on the two groups of pronouns, the *F*-ratio was much smaller (2.29), as expected, but it was still significant ($p \leq 0.001$), which was unexpected.

[7] With categorical data, it is not meaningful to ask what unit of distance is being used. The distance is only meaningful in relative terms. Thus 20.23 is half as far as 40.46 in **this particular data set**, but there is no independent standard against which 20.23 can be measured.

4. CONCLUSIONS

I used the new energy distance family of non-parametric statistics to evaluate competing claims about the category status of pronouns and determinatives. There is a 93.48% overlap between CGEL's categories and the *k*-means clusters (86.96% "adjusted agreement"). As expected, the results of the *k*-groups analysis strongly support not only the position that pronouns and determinatives are distinct groups, but that the particular words CGEL categorizes as pronouns and determinatives closely match the discovered clusters. This result suggests both that *k*-groups can be useful for discovering lexical categories and that CGEL's categories are strongly motivated by the features of the word forms.

The structure of the dendrogram provides further support for CGEL's categories. As expected, much of the structure reflects subcategories described in CGEL, though there are some unexpected elements too, such as the location of the temporal pronouns.

The results of the DISCO test show, as expected, that the two CGEL categories are significantly different ($p \leq 0.001$). Unexpectedly, the *F*-ratio of the two halves of the pronouns group, while smaller, is also significant ($p \leq 0.001$), meaning that the DISCO test can't be used as a simple way to assess category status.

The significant result with the two halves of the pronouns group may be understood to some degree by observing the dendrogram in Figure 1. The length of the branches from left to right reflects their distance in the high-dimensional matrix (R CORE TEAM, 2019). The leftmost pair of branches are the longest, and mostly reflect the split between determinatives and pronouns, but there is a good deal of structure **inside** each of those categories. The alphabetical order also imposed some structure with, for instance, all the interrogative pronouns in the second half of the list.[8]

Taken together these results appear to be consistent with the analysis set out in CGEL. They are not, however, entirely inconsistent with claims, such as HUDSON's (2004), that determinatives are a subcategory of pronouns or vice versa. More research is needed to see how these analyses perform with other categories (e.g., verbs & auxiliary verbs; nouns, pronouns, and determinatives; nouns & verbs; etc.) before rejecting or accepting such claims.

Having found statistical support for CGEL's analysis, the next step is to see whether it can be improved upon. Several possibilities for recategorization suggest themselves. CGEL categorizes the demonstratives (*this*, *that*, *these*, & *those*) with the determinatives, but calls them "borderline cases" (p. 422). It would be interesting to run the analysis with demonstratives categorized as pronouns and compare the *F*-ratios.

[8] I did run a third DISCO test on two completely randomized lists, and the result was not significant.

CGEL proposes two relative *which* words, a determinative (*during which time…*) and a pronoun (*the time in which we live*). The justification for this is that the pronoun has non-personal gender (it doesn't apply to persons, e.g., **the person which was there*) and that gender is a feature of pronouns. But CGEL's determinatives *you* and *us* (*us linguists*) have personal gender too, so this seems poorly motivated.

A third possibility is to try recategorizing the reciprocal words *each other* and *one another*, which CGEL has as pronouns based on the observation that they don't take any dependents. Their morphology, though, is much more like the compound determinatives (e.g., *anything* & *nobody*), suggesting they may fit better into that group.

A final example is that, in both the *k*-groups and the dendrogram, the so-called dummy pronouns were grouped with the determinatives. It could be useful to discover why this is the case.

Another interesting question that may be amenable to an approach like that presented here is discovering which of the 232 features used (or others that were overlooked) will turn out to be the most reliably predictive of a given category. Conversely, we may be able to identify members of the determinatives and pronouns that are most centrally located within their respective categories. It would be interesting to discover whether *the*, for instance, is really the most prototypical determinative or whether *it* is a more central pronoun than *you*.

The use of energy statistics can also be extended to other categories and to other languages, though there are considerable hurdles to applying this to other categories. Two stand out. The first is that the number of pronouns and determinatives is relatively small. Any attempt to compare determinatives and adjectives, for example, would presumably have to come up with a defensible sampling procedure for the adjectives. The second is that coding the semantic features of adjectives is likely much more complex than coding the features of the pronouns and determinative. Similar problems would apply to other open categories, though prepositions may be easier. Nevertheless, an existing scheme such as the UCREL Semantic Analysis System (RAYSON *et al.*, 2004) could be adapted to the problem.

A final general but important point is that the results of this new approach undermine Croft's (2001) broad accusations of language-internal methodological opportunism and his claim that "if one does choose one construction (or subset of constructions) to define a category, then one still has not accounted for the anomalous distribution pattern of the constructions that have been left out" (p. 41). The results from the three analyses performed here are consistent with CGEL's framework without the use of definitional features. In fact, this is what we would hope to find. A category defined perfectly by a single feature correlating with no other is of little value. One that has a cluster of generally related features is much more useful, even without any single perfectly reliable criterion.

It seems, then, that, when working in a grammar such as that presented in CGEL, with attention to a wide range of features and careful consideration of many cases, it is possible to discover useful categories that stand up to statistical scrutiny. Of course, this doesn't mean that methodological opportunism doesn't happen, or even that it is not the rule. But it does suggest that linguists can discover and describe lexical categories without being methodological opportunists.

5. ACKNOWLEDGEMENTS

I'd like to thank my family for their indulgence as I did the research and writing in my free time without funding. A version of this was presented at Linguistweets 2020. I'd like to thank Matthew Reynolds for help with R, Maria Rizzo for advice on DISCO and the Energy package for R, Bernhard Wälchli for pointing me to some very useful sources, and Aileen Bach and Tiago Luís Delgado for proofreading. Finally, I'd like to thank Dick Hudson for some helpful insights. Of course, the conclusions I draw are not necessarily endorsed by any of these kind folks.

REFERENCES

ABNEY, S. P. *The English noun phrase in its sentential aspect*. Massachusetts Institute of Technology, 1987.

CROFT, W. A. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford: Oxford University Press, 2001.

CRYSTAL, D. English. *Lingua*, v. 17, n. 1–2, p. 24–56. 1967.

EVERITT, B. S. *The Cambridge dictionary of statistics. 3. ed*. Cambridge: Cambridge University Press, 2006.

HASPELMATH, M. How comparative concepts and descriptive linguistic categories are different. *Aspects of Linguistic Variation*. De Gruyter Mouton, 2018. p. 83–114.

HUDDLESTON, R.; PULLUM, G. K. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press, 2002.

HUDSON, R. A. Are determiners heads? *Functions of Language*, v. 11, n. 1, p. 7–42, 2004.

HUDSON, R. A. *Language networks: The new word grammar*. Oxford: Oxford University Press, 2007.

JAMES, G. *et al. An introduction to statistical learning with applications in R*. 8th printing. New York: Springer, 2017.

LI, S. *K-groups: A generalization of k-means by energy distance*. Bowling Green State University, 2015. Disponível em: <https://etd.ohiolink.edu/apexprod/rws_olink/r/1501/10?p10_etd_subid=102410>.

MATTHEWS, P. H. *The positions of adjectives in English*. Oxford: Oxford University Press, 2014.

MESSICK, S. Standards of Validity and the Validity of Standards in Performance Assessment. *Educational Measurement: Issues and Practice*, v. 14, n. 4, p. 5–8, 1995.

NAU, N. *Wortarten und Pronomina: Studien zur lettischen Grammatik*. Wydział Neofilologii UAM w Poznaniu, 2016.

POSTAL, P. M. On so-called "pronouns" in English. In: KAYNE, R.; LEU, T.; ZANUTTINI, R. (Org.). *An annotated syntax reader: Lasting insights and questions*. Blackwell Publishing Ltd (originally published in 1966), 2014. p. 12–25.

R CORE TEAM. *R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing*, 2019. Disponível em: <https://www.r-project.org/>.

RAYSON, P. et al. The UCREL semantic analysis system. *Proceedings of the beyond named entity recognition semantic labelling for NLP tasks workshop*, p. 7–12, 2004. Disponível em: <http://eprints.lancs.ac.uk/1783/>.

REYNOLDS, B. Full matrix of English determinative and pronoun features. *Lingbuzz*, 2021. Disponível em: <https://ling.auf.net/lingbuzz/005747>.

RIZZO, M. L.; SZÉKELY, G. J. DISCO analysis: A nonparametric extension of analysis of variance. *Annals of Applied Statistics*, v. 4, n. 2, p. 1034–1055, 2010.

RIZZO, M. L.; SZÉKELY, G. J. Energy distance. *Wiley Interdisciplinary Reviews: Computational Statistics*, v. 8, n. 1, p. 27–38, 2016.

RIZZO, M. L.; SZÉKELY, G. J. *Package 'energy': E-statistics: Multivariate inference via the energy of data*, 2019. Disponível em: <https://github.com/mariarizzo/energy>

SAG, I. A.; WASOW, T.; BENDER, E. M. *Syntactic theory: A formal introduction. 2. ed*. Stanford, CA: Centre for the Study of Language and Information, 2003.

SOMMERSTEIN, A. H. On the so-called definite article in English. *Linguistic Inquiry*, v. 3, n. 2, p. 197–209, 1972.

APPENDIX: R CODE

```
### load packages ###
library(readr)
library(energy) #(RIZZO; SZÉKELY, 2019)
library(ape)
library(dplyr)


### Clear environment ###
rm(list = ls())


### Preparing the data ###
# load data WordListM as tibble with all columns as characters. The file has 73 determinatives, 65 pronouns
# this requires the comma-separated-values file named 73-65full.csv to be located in the default data folder
for R.
WordListM <- read_csv("73-65full.csv")

# convert all columns to factors
WordListM <- WordListM %>% mutate_if(is.character,as.factor)

# create WordListM as data frame from tibble
WordListM <- as.data.frame(WordListM)

# convert first column to rowname
rownames(WordListM) <- WordListM[, 1]
WordListM <- WordListM[, -1]

# convert data frame to matrix
WordListM <- data.matrix(WordListM, rownames.force = NA)


### k-groups ###
# run k-groups and display output
kgroups(WordListM [,-1], 2, iter.max = 10, nstart = 1, cluster = NULL)

# show list of k-groups cluster assignments
fitted(kgroups(WordListM[,-1], 2, iter.max = 10, nstart = 1, cluster = NULL))


### dendrogram ###
# create clusters
hc <- energy.hclust(dist(WordListM))

# plot clusters as dendrogram
plot(as.phylo(hc), cex = 0.3, label.offset = 0.1)


### DISCO ###
#Run DISCO for 73 determinatives & 65 pronouns
eqdist.etest(WordListM, sizes=c(73, 65), R=999, method="discoF")

#Rerun DISCO on 73 determinatives split in half to test sensitivity of discoF
eqdist.etest(WordListM[c(1:73),], sizes=c(35,38), R=999, method="discoF")
```