Multifactorial Exploratory Approaches: fundamentals

Guillaume Desagulier

Multifactorial Exploratory Approaches

fundamentals

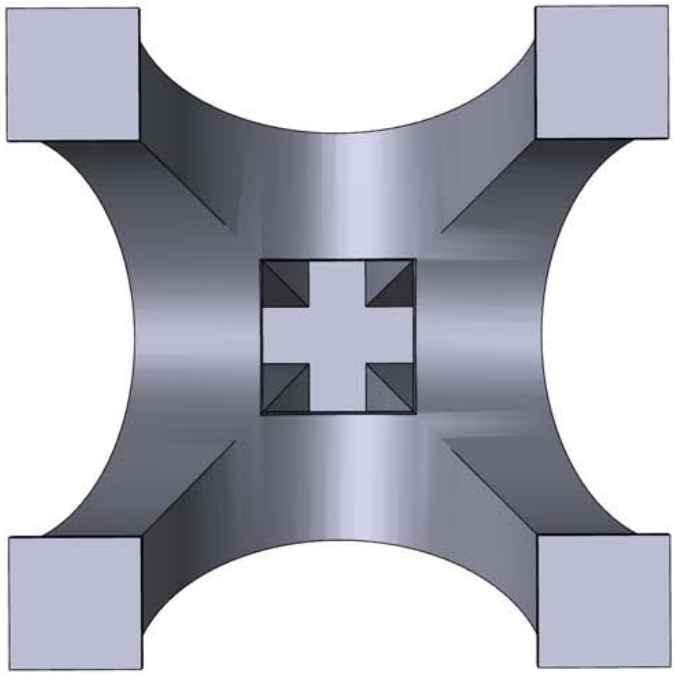
Guillaume Desagulier[1]

[1]MoDyCo (UMR 7114)
Paris 8, CNRS, Paris Nanterre
Institut Universitaire de France
gdesagulier@univ-paris8.fr

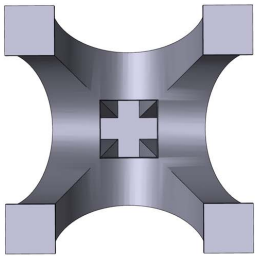
Corpus Linguistics Summer School 2019
June 24th, 2019
University of Birmingham

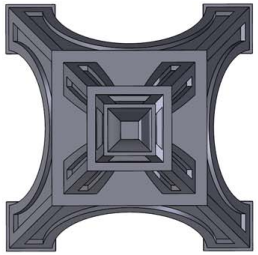
outline

an example from home

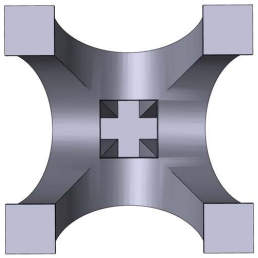


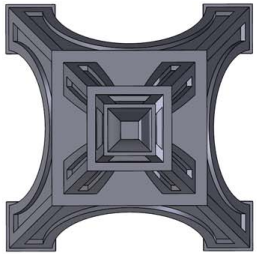
rotation

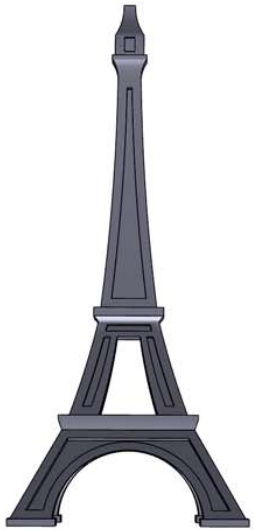




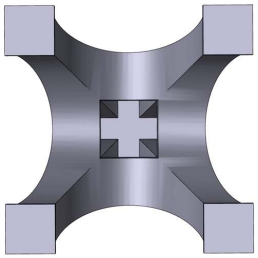
rotation

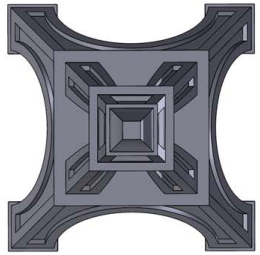


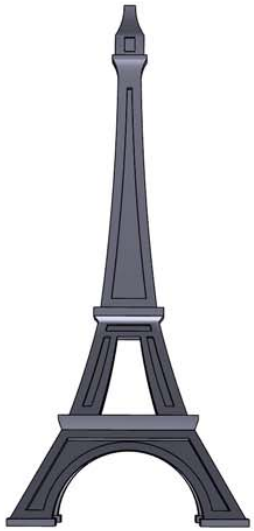




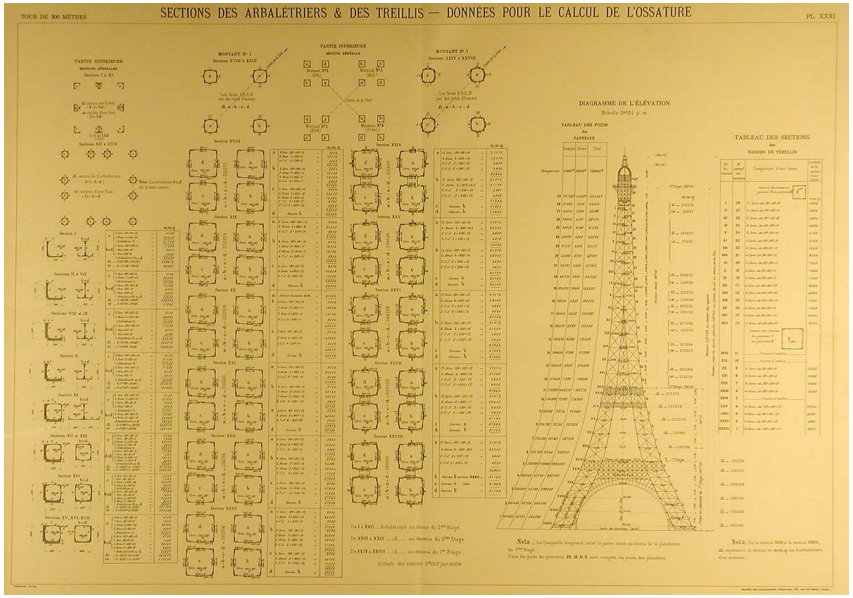
rotation







reduction


SECTIONS DES ARBALÉTRIERS & DES TREILLIS — DONNÉES POUR LE CALCUL DE L'OSSATURE
TOUR DE 300 MÈTRES
PL. XXXI
DIAGRAMME DE L'ÉLÉVATION
TABLEAU DES SECTIONS

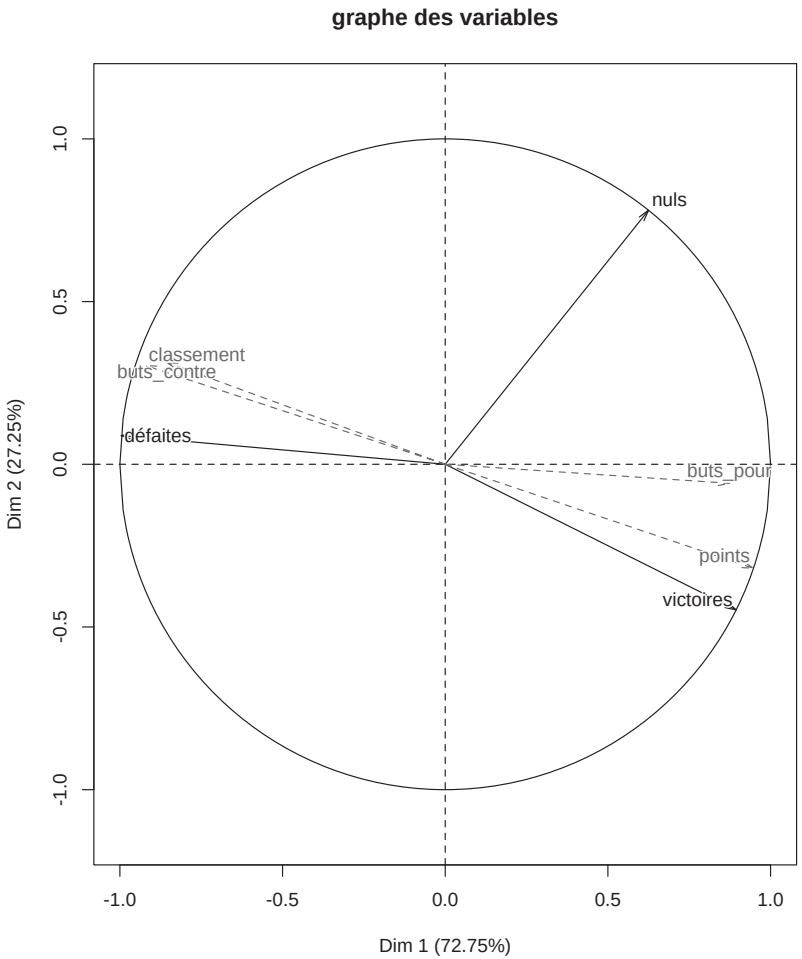
a more quantitative example

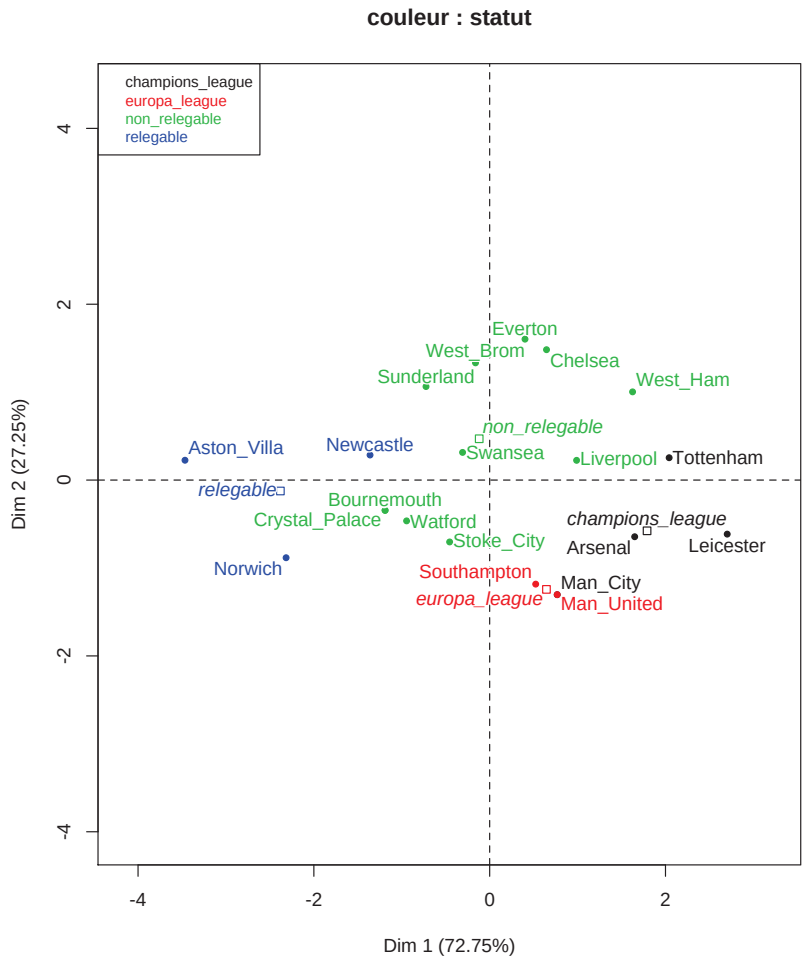
Premier League 2015–2016, final table

team	ranking	points	wins	draws	losses	goals for	goals against	status
Leicester	1	81	23	12	3	68	36	champions league
Arsenal	2	71	20	11	7	65	36	champions league
Tottenham	3	70	19	13	6	69	35	champions league
Man City	4	66	19	9	10	71	41	champions league
Man United	5	66	19	9	10	49	35	europa league
Southampton	6	63	18	9	11	59	41	europa league
West Ham	7	62	16	14	8	65	51	safe
Liverpool	8	60	16	12	10	63	50	safe
Stoke City	9	51	14	9	15	41	55	safe
Chelsea	10	50	12	14	12	59	53	safe
Everton	11	47	11	14	13	59	55	safe
Swansea	12	47	12	11	15	42	52	safe
Watford	13	45	12	9	17	40	50	safe
West Brom	14	43	10	13	15	34	48	safe
Crystal Palace	15	42	11	9	18	39	51	safe
Bournemouth	16	42	11	9	18	45	67	safe
Sunderland	17	39	9	12	17	48	62	safe
Newcastle	18	37	9	10	19	44	65	relagation zone
Norwich	19	34	9	7	22	39	67	relagation zone
Aston Villa	20	17	3	8	27	27	76	relagation zone

a more quantitative example

Premier League 2015–2016, visualisation





looking for patterns in the data

Once corpus linguists have collected sizeable amounts of observations, they look for patterns in the data. When the data set is too large, it becomes impossible to summarize the table with the naked eye and summary statistics are needed. This is where exploratory data analysis steps in.

exploring a data set

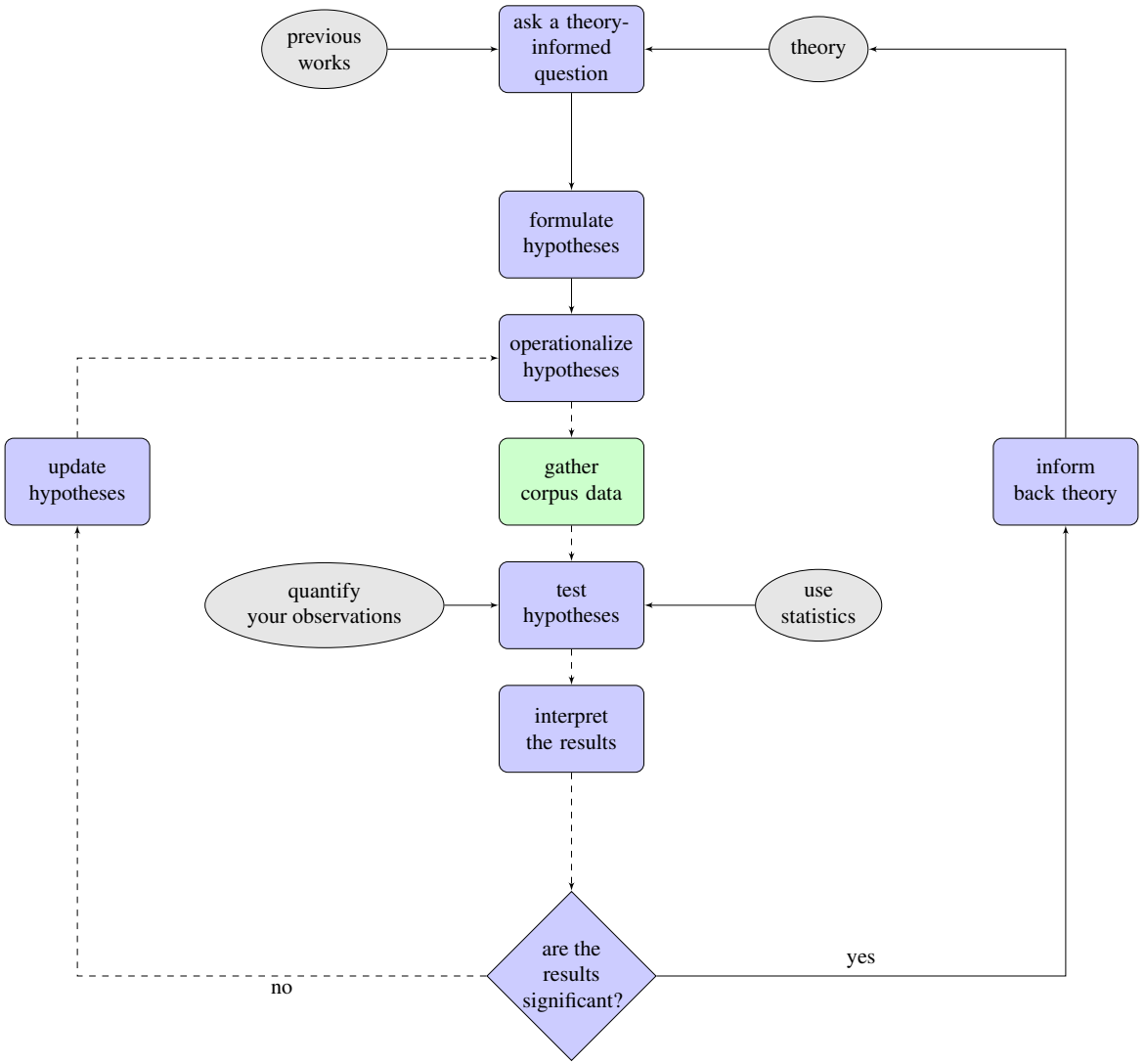
Exploring a data set means separating meaningful trends from the noise (i.e. "random" distributions)[1]

[1] Even though "language is never, ever, ever random" (Kilgarriff 2005)

generating hypotheses

In theory, exploratory data analysis is used to generate hypotheses because the linguist does not yet have any assumption as to what kinds of trends should appear in the data.
In practice, however, linguists collect observations in the light of specific variables precisely because they expect that the latter influence the distribution of the former.

the empirical cycle



multifactorial, multivariate, and multidimensional

When a linguistic phenomenon is influenced by several factors at the same time, its analysis is multifactorial. Experience tells us that, arguably, just about anything in language is multifactorial.

Bresnan et al. (2007)

The dative alternation in English is influenced by several factors, such as:

- the meaning of the verb
- the length/animacy/definiteness/pronominality/accessibility of the recipient/theme
- the realization of the recipient
- etc.

multifactorial, multivariate, and multidimensional

Once operationalized by the linguist, these multiple factors are captured by means of several independent variables. When observations of the linguistic phenomenon are captured by several variables, the analysis is multivariate.

Bresnan et al. (2007)

Each of the 3263 observations in the `dative` data set is described by 15 variables

```
> install.packages("languageR")
> library(languageR)
> str(dative)
```

terminology

multifactorial, multivariate, and multidimensional

We are now entering the world of data tables from a maths/stats viewpoint! The analysis becomes (multi)dimensional when the complex table is decomposed into meaningful dimensions.
The dimensions can be:

- explicit
- implicit

terminology

explicit dimensions

Table 1: a word-word cooccurrence matrix

	different	quite	thing	writing	natural	place	sort	become	men	...
different	535	491	43	3	0	8	21	2	5	...
quite	491	3048	102	4	17	37	30	19	23	...
thing	43	102	176	1	1	0	5	2	0	...
writing	3	4	1	11	0	0	0	0	0	...
natural	0	17	1	0	24	1	1	0	0	...
place	8	37	0	0	1	88	2	0	0	...
sort	21	30	5	0	1	2	75	0	2	...
become	2	19	2	0	0	0	0	36	1	...
men	5	23	0	0	0	0	2	1	38	...
...	...	...	...	...	...	...	...	...	...	...

terminology

implicit dimensions

Table 2: a word-vector matrix of some adjectives from the BNC (snapshot)

adjectives	V1	V2	V3	V4	V5	V6	...
gloomy	-0.36405	-0.44487	-0.33327	-0.16695	-0.52404	0.31066	...
sacred	0.60337	-0.20526	-0.042822	-0.33008	-0.68957	0.26654	...
jaundiced	-0.32168	-0.58319	-0.34614	-0.12474	0.10368	0.1733	...
loud	0.24615	-0.24904	-0.18212	-0.14834	-0.06532	-0.3393	...
memorable	0.30206	0.20307	0.062304	0.66816	0.048326	0.034361	...
justified	-0.080959	-0.23694	-0.43372	-0.31442	-0.31528	0.0057226	...
scant	-0.14467	-0.29329	0.10832	-0.11123	-0.57925	-0.27022	...
continuous	-0.15253	-0.082764	-0.40871	-0.53719	0.0822	-0.31482	...
imposing	0.32043	0.155	-0.10547	-0.23157	-0.35657	-0.097553	...
weighty	0.085281	0.015087	0.58454	0.0094917	-0.082617	0.36811	...
...	...	...	...	...	...	...	...

multifactorial exploratory methods

- multifactorial exploratory methods are used to summarize such complex tables
- the tables are complex objects of which we want to get a synthetic view.
- we do it by finding clusters in the data. This is no trivial task because we need to make sure the clusters are valid.
- the reward is a graphic representation of the data set in terms of neat clusters.

5 exploratory methods

the first four are based on eigenvalue decomposition

- correspondence analysis (CA)
- multiple correspondence analysis (MCA)
- principal component analysis (PCA)
- exploratory factor analysis (EFA)

the fifth one is more recent

- t-SNE

you do not compile a data set randomly

what does *exploratory* mean? (again)

the linguist makes no assumption as to what kinds of groupings are to be found in the data. In practice, however, you compile a table of data because you expect to find meaningful groupings. Therefore, if you find no meaningful grouping, this is because your rows and your columns are independent. Chances are that you might want to rethink the design of your study, especially your choice of explanatory variables.

goal

- we seek to explore a cloud of points from a data set in the form of a *rows* $\times$ *columns* table with as many dimensions as there are columns.
- like a complex object in real life, a data table has to be rotated so as to be observed from an optimal angle

dimensions

- although the dimensions of a data table are eventually projected in a two-dimensional plane, they are not spatial dimensions
- if the table has K columns, the data points are initially positioned in a space $\mathbb{R}$ of K dimensions

dimensionality reduction

- to allow for easier interpretation, dimensionality-reduction methods decompose the cloud into a smaller number of meaningful planes.
- the methods covered in this course summarize the table by measuring how much variance there is and decomposing the variance into proportions.
- these proportions are eigenvalues in CA, MCA, and PCA. They are loadings in EFA (and a special kind of PCA not covered in this course).[2]

[2] See **baayen2008analyzing**.

graphic summary

- all four methods offer graphs that facilitate the interpretation of the results
- although convenient, these graphs do not replace a careful interpretation of the numerical results.

differences

The main difference between these methods pertain mainly to the kind of data that one works with

CA

- CA takes as input a contingency table, i.e. a table that cross-classifies observations on a number of categorical variables
- entries in each cell are integers, namely the number of times that observations (in the rows) are seen in the context of the variables (in the columns).

CA

Table 3 is an example of a contingency table. It displays the frequency counts of four types of nouns (rows) across three corpus files from the BNC-XML (columns).

Table 3: An example of a contingency table

	A1J.xml	A1K.xml	A1L.xml	row totals
NN0	136	14	8	158
NN1	2236	354	263	2853
NN2	952	87	139	1178
NP0	723	117	71	911
column totals	4047	572	481	5100

MCA

- MCA takes as input a case-by-variable table such as Table 4.
- the table consists of *i* individuals or observations (rows) and *j* variables (columns).

Table 4: A sample input table for MCA (Desagulier 2017, p. 36)

corpus file	mode	genre	exact match	intensifier	syntax	adjective
KBF.xml	spoken	conv	*a quite ferocious mess*	quite	preadjectival	*ferocious*
AT1.xml	written	biography	*quite a flirty person*	quite	predeterminer	*flirty*
A7F.xml	written	misc	*a rather anonymous name*	rather	preadjectival	*anonymous*
ECD.xml	written	commerce	*a rather precarious foothold*	rather	preadjectival	*precarious*
B2E.xml	written	biography	*quite a restless night*	quite	predeterminer	*restless*
AM4.xml	written	misc	*a rather different turn*	rather	preadjectival	*different*
F85.xml	spoken	unclassified	*a rather younger age*	rather	preadjectival	*younger*
J3X.xml	spoken	unclassified	*quite a long time*	quite	predeterminer	*long*
KBK.xml	spoken	conv	*quite a leading light*	quite	predeterminer	*leading*

MCA

- historically, MCA was developed to explore the structure of surveys in which informants are asked to select an answer from a list of suggestions.
- for example, the question "According to you, which of these disciplines best describe the hard sciences: physics, biology, mathematics, computer science, or statistics?" requires informants to select one category.

PCA

- PCA takes as input a table of data of *i* individuals or observations (rows) and *j* variables (columns)
- the method handles continuous and nominal data
- the continuous data may consist of means, reaction times, formant frequencies, etc.
- the categorical/nominal data are used to tag the observations

PCA

Table 6 is a table of 6 kinds of mean frequency counts further described by 3 kinds of nominal information.

Table 5: A sample data frame (Lacheret-Dujour et al. 2019)

corpus sample	fPauses	fOverlaps	fFiller	fProm	fPI	fPA	subgenre	interactivity	planning type
D0001	0.26	0.12	0.14	1.79	0.28	1.54	argumentation	interactive	semi-spontaneous
D0002	0.42	0.11	0.10	1.80	0.33	1.75	argumentation	interactive	semi-spontaneous
D0003	0.35	0.10	0.03	1.93	0.34	1.76	description	semi-interactive	spontaneous
D0004	0.28	0.11	0.12	2.29	0.30	1.79	description	interactive	semi-spontaneous
D0005	0.29	0.07	0.23	1.91	0.22	1.69	description	semi-interactive	spontaneous
D0006	0.47	0.05	0.26	1.86	0.44	1.94	argumentation	interactive	semi-spontaneous
...	...	...	...	...	...	...	...	...	...

EFA

Like PCA, EFA takes as input a table of continuous data. However, it does not commonly accommodate nominal data. Typically, Table 6 minus the 3 columns of nominal data can serve as input for EFA.

Table 6: A sample data frame (Lacheret-Dujour et al. 2019)

corpus sample	fPauses	fOverlaps	fFiller	fProm	fPI	fPA
D0001	0.26	0.12	0.14	1.79	0.28	1.54
D0002	0.42	0.11	0.10	1.80	0.33	1.75
D0003	0.35	0.10	0.03	1.93	0.34	1.76
D0004	0.28	0.11	0.12	2.29	0.30	1.79
D0005	0.29	0.07	0.23	1.91	0.22	1.69
D0006	0.47	0.05	0.26	1.86	0.44	1.94
...	...	...	...	...	...	...

exploring is not predicting

The methods presented in this course are exploratory, as opposed to explanatory or predictive. They help find structure in multivariate data thanks to observation groupings. The conclusions made with these methods are therefore valid for the corpus only.

exploring is not predicting

For example, we shall see that middle-class female speakers aged 25 to 59 display a preference for the use of *bloody* in the British National Corpus. This finding should not be extended to British English in general. Indeed, we may well observe different tendencies in another corpus of British English.

exploring is not predicting

Neither should the conclusions made with exploratory methods be used to make predictions. Of course, exploratory methods serve as the basis for the design of predictive modeling, which uses the values found in a sample to predict values for another sample.

exploring is not predicting

Glynn (2014)

Expanding on Gries (2006), Glynn (2014) finds that usage features and dictionary senses are correlated with dialect and register thanks to two exploratory multivariate techniques (correspondence analysis and multiple correspondence analysis). To confirm these findings, Glynn (ibid.) turns to logistic regression. This confirmatory multivariate technique allows him to specify which of the usage features and dictionary senses are significantly associated with either dialect or register, and determine the importance of the associations.

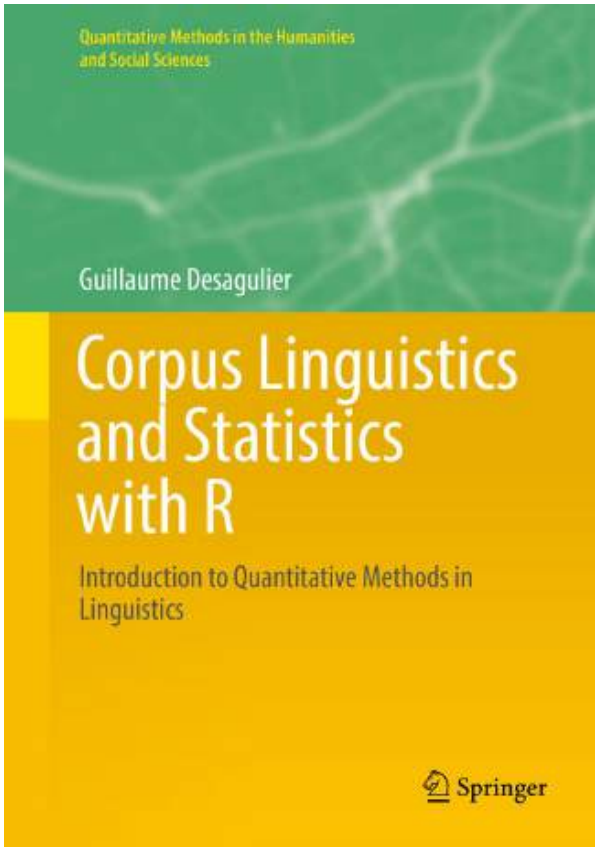
so why explore after all?

Nowadays, many linguists jump to powerful predictive methods (such as logistic regression or discriminant analysis) without going through the trouble of exploring their data sets first.
This is a shame because the point of running a multifactorial exploratory analysis is to generate fine research hypotheses, which the far more powerful predictive methods can only benefit from.

Around the word

https://corpling.hypotheses.org/

Corpus Linguistics and Statistics with R



chapter 10 – (Desagulier 2017)

Practical Handbook of Corpus Linguistics

Guillaume Desagulier (to appear). "Multifactorial exploratory approaches." In: *Practical Handbook of Corpus Linguistics*. Ed. by Magali Paquot and Stefan Thomas Gries. New York: Springer

Bibliography I

- Bresnan, Joan et al. (2007). "Predicting the dative alternation." In: *Cognitive Foundations of Interpretation*, pp. 69–94.
- Desagulier, Guillaume (to appear). "Multifactorial exploratory approaches." In: *Practical Handbook of Corpus Linguistics*. Ed. by Magali Paquot and Stefan Thomas Gries. New York: Springer.
- – (2017). "Clustering Methods." In: *Corpus Linguistics and Statistics with R*. New York, NY: Springer, pp. 239–294.
- Glynn, Dylan (2014). "The many uses of *run*." In: *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy*. Ed. by Dylan Glynn and Justyna A. Robinson. Vol. 43. Human Cognitive Processing. John Benjamins, pp. 117–144.
- Gries, Stefan Th (2006). "Corpus-based methods and cognitive semantics: The many senses of to run." In: *Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax and Lexis*. Ed. by Stefan Th Gries and Anatol Stefanowitsch. Mouton de Gruyter, pp. 57–99.

Bibliography II


Kilgarriff, Adam (2005). "Language is never, ever, ever, random." In: *Corpus Linguistics and Linguistic Theory* 1.2, pp. 263–276. URL: http://dx.doi.org/10.1515/cllt.2005.1.2.263.


Lacheret-Dujour, Anne et al. (2019). "The distribution of prosodic features in the Rhapsodie corpus." In: *Rhapsodie: A prosodic and syntactic treebank for spoken French*. Ed. by Anne Lacheret-Dujour and Sylvain Kahane. Studies in Corpus Linguistics 89. John Benjamins. Chap. 17, pp. 315–338.