А.А. Кожухарь

Научный руководитель:
М.А. Даниэль

КЛАСТЕРНЫЙ АНАЛИЗ В ТИПОЛОГИИ ПОРЯДКА СЛОВ

Введение

Цель исследования — разработать и описать системные корреляции по значениям параметров порядка слов в языках мира, представленных базой World Atlas of Language Structures (WALS), используя статистические методы кластеризации и уменьшения мерности данных.

1. Ограничения

Основное ограничение исследования — лакуны, т.е. отсутствие данных о категории в отдельном языке. Такие пустоты кодировались символом знака вопроса в составленной базе данных.

Лакуны представляют серьезную концептуальную проблему для статистического анализа, который работает с ограниченным алфавитом кодировки, следовательно, принимает одинаковый набор символов за один тип категории, тогда как лакуна не является отдельным значением категории. За обозначением лакуны может стоять любое ранее использованное значение (или принципиально новое).

В работе проблема лакун решается максимально тривиальным образом. Из рабочей выборки изымались языки с наибольшим количеством лакун таким образом, чтобы результирующее количество лакун стало статистически незначимым для масштабов выборки.

2. Материал

Использованные в работе языковые данные были взяты с портала WALS и в дальнейшем были стратифицированы по следующим параметрам: ареал, языковая семья и genus. Таким образом, рабочая языковая выборка состоит из 520 языков (из 2679 доступных на портале WALS), которые распределены по шести макроареалам. Выборка составлена на основе материала 120 языковых семей (из доступных на портале WALS 215 языковых семей), в ней задействованы 290 genera (из 529 представленных на портале WALS).

2.1. База данных

База данных была представлена в виде таблицы сопряженности, где по горизонтали расположены языки, а по вертикали — категории порядка слов или значения категорий. Таблица по категориям содержит следующие категории: порядок предиката и прямого объекта, порядок субъекта и предиката, положение адлога, положение указательного местоимения, порядок числительного и имени, порядок релятивной клаузы и вершины, порядок адъективного модификатора и прилагательного, положение вопросительной частицы, положение адвербиального субординатора и клаузы, положение отрицания относительно предиката.

Таблица со значениями категорий содержит 37 колонок. Названия колонок соответствуют порядку составляющих, которые допускает данная категория.

Информация на пересечении варьируется в зависимости от типа колонки: если в названии колонки указана категория, то на пересечении будет указан порядок составляющих, характерный для этого языка. В колонке со значением категории на пересечениях могут располагаться только «0» и «1»: соответственно «1» — если категория есть в языке, «0» — если категория в языке отсутствует.

Дополнительно в этих базах данных на пересечении может использоваться «?» для обозначения лакуны.

3. Методика

Для получения желаемых результатов использовались следующие статистические методы:

1) Multiple Correspondence Analysis (далее — MCA) — метод множественных соответствий для кластеризации параметров;

2) парная и частная корреляции для формулировки универсалий.

3.1. Статистический анализ

Статистический анализ состоял из двух частей. На первом этапе было выяснено, каким образом происходит кластеризация параметров. На втором этапе исчислялось расстояние между категориями в кластере.

3.1.1. Кластеризация

Для построения кластеров использовался MCA, который уменьшает мерность данных с максимальной экономией извлеченной информации.

Интерпретация графиков MCA осуществляется по степени сближенности точек между собой [Abdi, Valentin, 2007, p. 8]. Если две разные категориальные переменные расположены близко друг к другу на графике, значит, они показывают одинаковое распределение разных значений категорий.

3.1.2. Корреляция

Поскольку MCA не предоставляет численного результата, вторым этапом работы является расчет степени скоррелированности каждой категории. Для этого используется подсчет попарной и частной корреляций между значениями категорий.

Для подсчета численной зависимости одного значения категории от другого использовалась попарная корреляция, для подсчета зависимости одной категории от другой — коэффициент Крамера (Cramer's V).

Расчет парной и частной корреляции проводился на части рабочей выборки, состоящей из 144 языков, которые не продемонстрировали лакун.

4. Результаты

4.1. Multiple correspondence analysis

MCA предоставляет возможность кластеризовать не только значения категорий, но и сами категории (см. рис. 1).

Очевидную корреляцию демонстрируют категории Adpo (положение адпозиции относительно модифицируемого имени) и V.O (положение переходного предиката и прямого объекта). Следующей по степени скоррелированности является пара AdvSub.Clause (положение адвербиального субординатора относительно клаузы, которую он модифицирует) и Gen.N (положение имени и генитивного аргумента). Последней максимально скоррелированной парой категорий является Adj.N (положение прилагательного и имени) и Deg.Adj (положение адъективного модификатора и модифицируемого прилагательного). Прочие пары или тройки категорий расположены сравни-

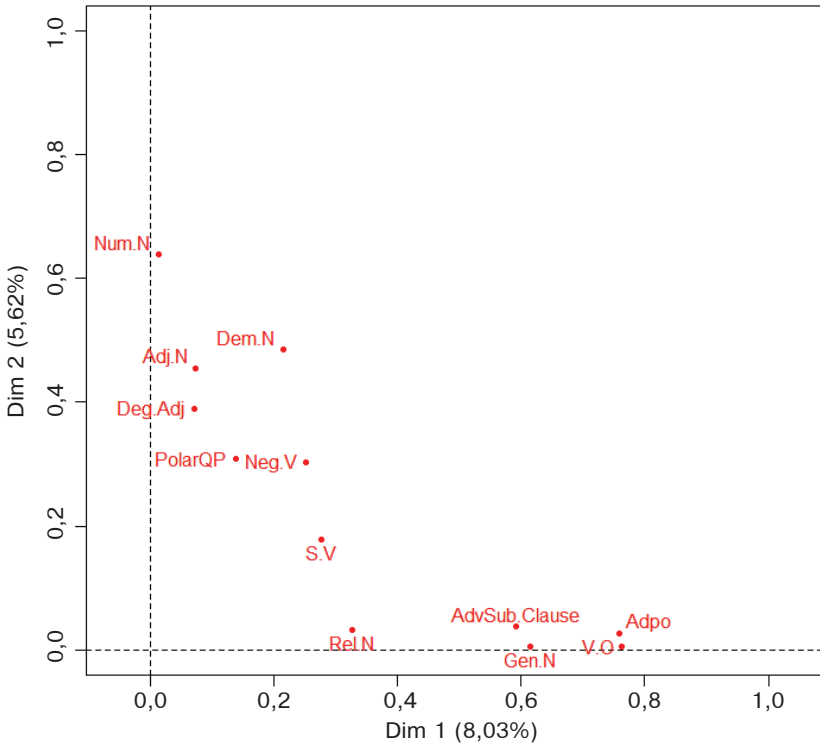


Рис. 1. Кластеризация по категориям порядка слов

тельно далеко друг от друга на графике, поэтому следующим шагом должен стать анализ не категорий, а значений категорий.

Следующие графики (см. рис. 2, 3) демонстрируют скоррелированность значений параметров.

На рис. 2 показано шесть областей, где категории находятся в сравнительной близости друг к другу. Первый кластер содержит значения категорий: Adj.N_AdjN, Deg.Adj_DegAdj, PolarQP_SecPos. Второй кластер включает значения Adpo_InPo, Neg.V_(NegV), NDO(VO) (см. рис. 3). Третий кластер содержит значения Adj.N_NAdj, AdvSub. Clause_InternSub. Четвертый и пятый кластеры охватывают значения V.O_VO, AdPo_PrN и V.O_OV, AdPo_NPo соответственно. Шестой кластер содержит значения PolarQP_TwoPos, Neg.V_NegV. Прочие значения категорий находятся в сравнительной отдаленности друг от друга.

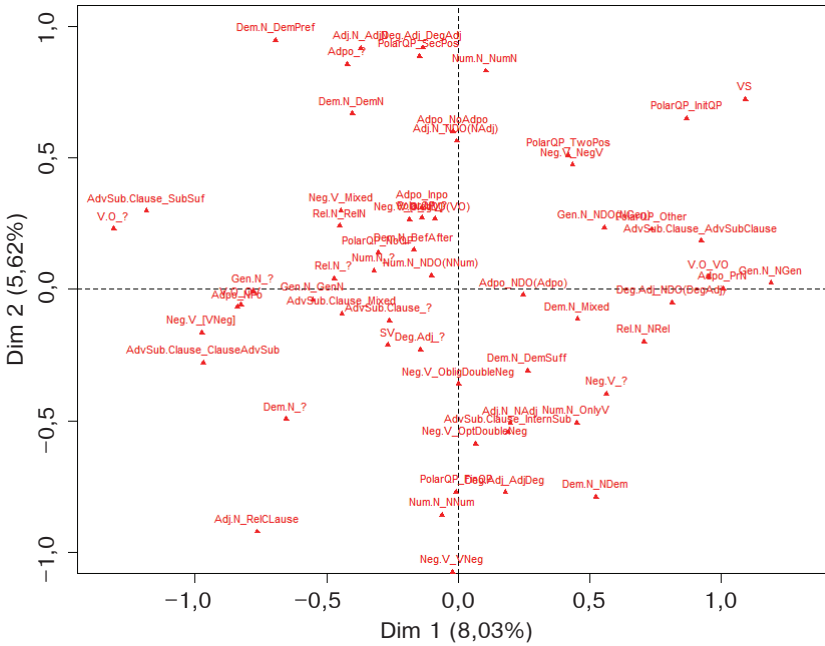


Рис. 2. Кластеризация по значениям категорий порядка слов

4.2. Cramer's V

В табл. 1, 2 представлены числовые значения коэффициента Крамера, которые свидетельствуют соответственно о сильной и средней связи параметров.

Таблица 1

Значения по результатам коэффициента Крамера, свидетельствующие о сильной связи параметров

Значения	Cramer's V
~ VO + PrN	0,538
~ OV + PrN	0,507
~ VO + NPo	0,555
~ OV + NPo	0,538

Результаты MCA и применения коэффициента Крамера дают сопоставимые параметры.

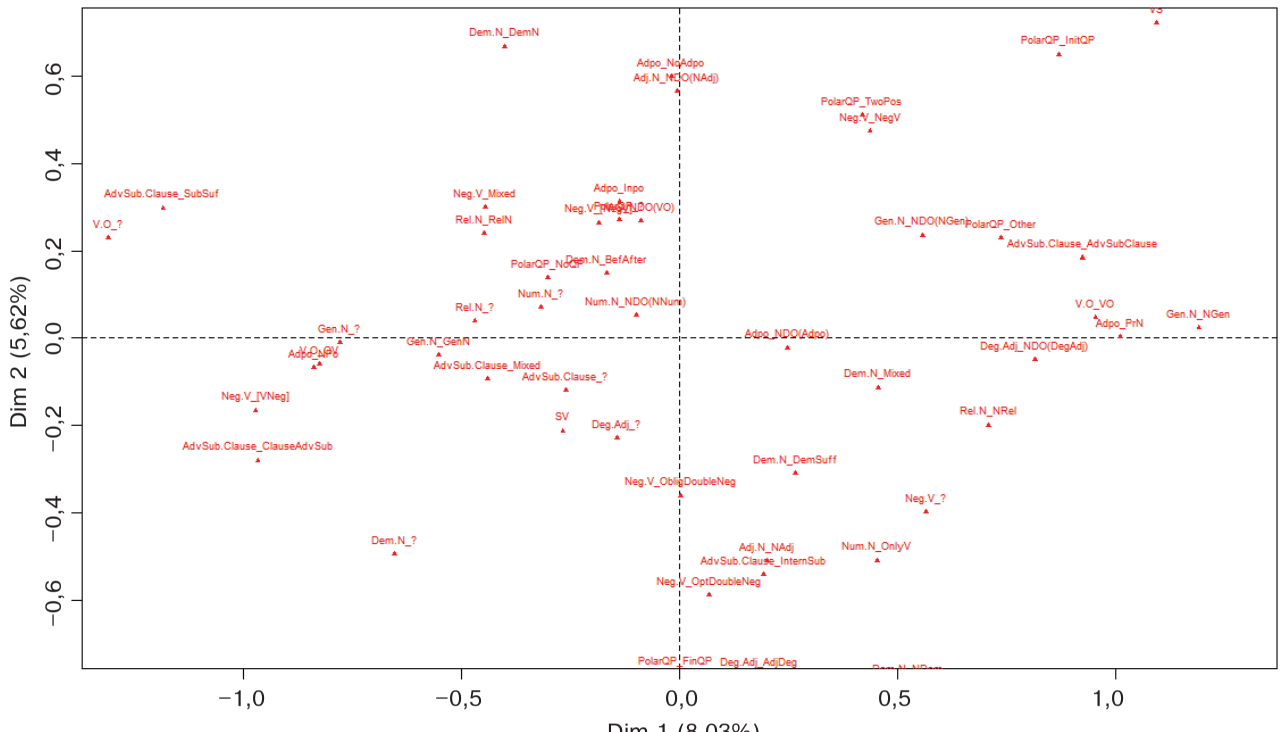


Рис. 3. Кластеризация по значениям категорий порядка слов (максимальный масштаб)

Таблица 2

Значения по результатам коэффициента Крамера, свидетельствующие о средней связи параметров

Значения	Cramer's V	Значения	Cramer's V	Значения	Cramer's V
~ VO + SV	0,461	~ SV + GenN	0,363	~ SV + NNum	0,315
~ VO + VS	0,45	~ SV + NGen	0,381	~ DemN + DegAdj	0,302
~ OV + SV	0,471	~ VS + GenN	0,355	~ VO + RelN	0,316
~ OV + VS	0,428	~ VS + NGen	0,374	~ VO + NRel	0,316
~ SV + PrN	0,398	~ PrN + GenN	0,463	~ OV + RelN	0,335
~ SV + NPo	0,408	~ PrN + NGen	0,448	~ OV + NRel	0,335
~ VS + PrN	0,352	~ NPo + GenN	0,449	~ PrN + RelN	0,312
~ VS + NPo	0,369	~ NPo + NGen	0,402	~ PrN + NRel	0,312
~ VO + GenN	0,423	~ AdjN + DemN	0,339	~ NPo + RelN	0,302
~ VO + NGen	0,398	~ AdjN + NDem	0,323	~ NPo + NRel	0,302
~ OV +GenN	0,409	~ NAdj + DemN	0,353	~ VO + AdvSubClause	0,414
~ OV + NGen	0,38	~ NAdj + NDem	0,34	~ VO + ClauseAdvSub	0,3
~ DemN + NumN	0,324	~ NDem + NumN	0,314	~ OV + AdvSubClause	0,414
~ DemN + NNum	0,325	~ NDem + NNum	0,324	~ OV + ClauseAdvSub	0,312
~ NGen + AdvSubClause	0,36	~ NPo + AdvSubClause	0,403	~ SV + AdvSubClause	0,31
~ SV + NegV	0,329	~GenN + AdvSubClause	0,391	~ PrN + AdvSubClause	0,41

4.3. Частная и парная корреляции

Чтобы выделить действительно связанные параметры, использовался дополнительный тест, который считает корреляцию между значениями, фиксируя значения прочих величин, — частную корреляцию. Если полученное значение частной корреляции меньше значения парной корреляции, значит, на зависимость этих значений влияет третье значение [Baba et al., 2004].

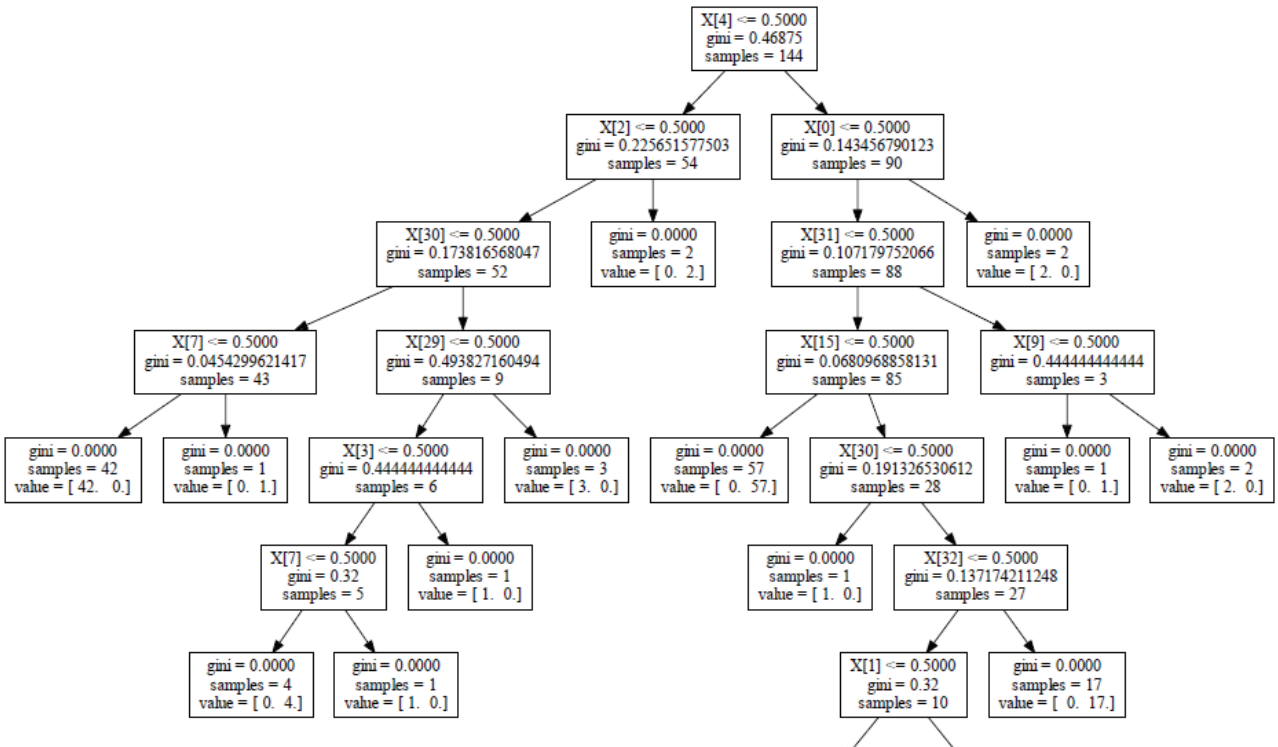
Самую сильную положительную корреляцию, как и ожидалось, показала пара значений VO + PrN и OV + NPo — 0,793 и 0,753 соответственно. Подсчет частной корреляции, последовательно исключающий каждый из рассмотренных признаков порядка слов (SV, VS, GenN, NGen, AdjN, NAdj, DemN, NDem, RelN, NRel, DegAdj, AdjDeg, InitQP, FinQP, AdvSubClause, ClauseAdvSub, NegV, VNeg, NNum, NumN), показал, что ни один из этих признаков не уменьшает коэффициент корреляции на статистически значимую разницу, т.е. связь этих двух параметров не обусловлена третьими факторами. Другой независимой парой признаков является AdjN + DemN и NDem + + NAdj, поскольку ни одно из третьих значений не уменьшает коэффициент корреляции на статистически значимую разницу. Пары значений VS + PrN и SV + NPo, SV + GenN и GenN + VS, PrN + NGen и NPo + GenN демонстрируют высокий коэффициент корреляции благодаря сильному влиянию пар значений VO + PrN и OV + NPo. Значения GenN, NGen, RelN, NRel, AdvSubClause и ClauseAdvSub, видимо, целесообразно объединять в группу значений, зависимых от соположения переходного предиката и прямого объекта и соположения адлога и имени.

Вторая группа зависимых значений привязана к парам AdjN + + DemN и NDem + NAdj и включает следующие значения категорий порядка слов: NumN, NNum, DegAdj, AdjDeg.

В соответствии с тестами на коэффициент частной корреляции ядерными порядками, от которых далее идет отсчет для порядка прочих элементов, являются порядок предиката и прямого объекта, порядок адлога и имени, порядок прилагательного и имени и порядок демонстратива и имени.

5. Решение проблемы лакун

Поскольку основной проблемой настоящего исследования автор считает сравнительно большое количество пустот в языковом материале, был предложен метод заполнения лакун в базе данных.


X[4] <= 0.5000
gini = 0.46875
samples = 144
X[2] <= 0.5000
gini = 0.225651577503
samples = 54
X[0] <= 0.5000
gini = 0.143456790123
samples = 90
X[30] <= 0.5000
gini = 0.173816568047
samples = 52
gini = 0.0000
samples = 2
value = [0. 2.]
X[31] <= 0.5000
gini = 0.107179752066
samples = 88
gini = 0.0000
samples = 2
value = [2. 0.]
X[7] <= 0.5000
gini = 0.0454299621417
samples = 43
X[29] <= 0.5000
gini = 0.493827160494
samples = 9
X[15] <= 0.5000
gini = 0.0680968858131
samples = 85
X[9] <= 0.5000
gini = 0.444444444444
samples = 3
gini = 0.0000
samples = 42
value = [42. 0.]
gini = 0.0000
samples = 1
value = [0. 1.]
X[3] <= 0.5000
gini = 0.444444444444
samples = 6
gini = 0.0000
samples = 3
value = [3. 0.]
gini = 0.0000
samples = 57
value = [0. 57.]
X[30] <= 0.5000
gini = 0.191326530612
samples = 28
gini = 0.0000
samples = 1
value = [1. 0.]
gini = 0.0000
samples = 2
value = [2. 0.]
X[7] <= 0.5000
gini = 0.32
samples = 5
gini = 0.0000
samples = 1
value = [1. 0.]
gini = 0.0000
samples = 1
value = [1. 0.]
X[32] <= 0.5000
gini = 0.137174211248
samples = 27
gini = 0.0000
samples = 4
value = [0. 4.]
gini = 0.0000
samples = 1
value = [1. 0.]
X[1] <= 0.5000
gini = 0.32
samples = 10
gini = 0.0000
samples = 17
value = [0. 17.]

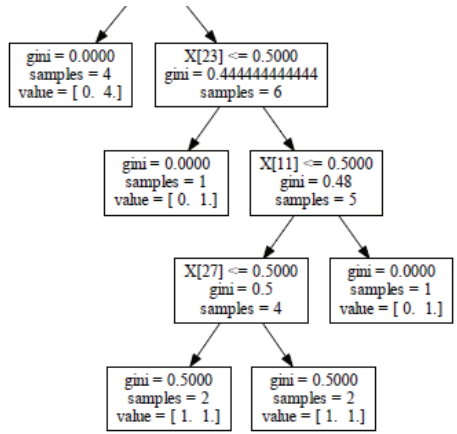


Рис. 4. Дерево решений, реализующее присвоение порядка VO или OV языку[2]

[2] Значение Gini (Gini impurity) — это мера того, как часто случайно выбранному элементу из набора будет присвоен неверный элемент в соответствии с правилами присвоения, описанными выше. Значение gini минимально (равно нулю), когда всем рассматриваемым случаям может быть присвоено только одно таргетное значение. Чем больше значение gini, тем выше вероятность ошибки при присвоении значения категории в конкретном узле. В случае рис. 4 таргетными значениями являются соответственно значения VO и OV.

Решение основывается на машинном обучении. Программа получает на вход набор наблюдений и формулирует правила распределения нулей и единиц для категории порядка слов в базе данных, опираясь на имеющиеся данные о наблюдениях, т.е. на нули и единицы у других значений категорий порядка слов. Выведенные правила программа экстраполирует на другие языки рабочей выборки. Имеющиеся данные сравнивались с полученными с помощью алгоритма машинного обучения. На первом этапе запуска программы предсказание оказывалось верным на 86%, на последующих этапах запуска правильность предсказания доходила до 90%. Для реализации машинного обучения на Python использовался пакет sklearn[1].

Машинное обучение в данном случае не только с высокой вероятностью предсказывает значение категории на месте лакуны, но и формулирует правила для предсказания, которые, будучи представленными в виде дерева решений (decision tree), могут быть использованы для формулировки лингвистических универсалий.

Дополнительно предлагается использовать метод random forest («случайный лес») поскольку вместо одного дерева решений, как в простом машинном обучении, метод выстраивает «ансамбль» деревьев решений.

Результатом работы программы является дерево решений, демонстрирующее алгоритм выбора категории (см. рис. 4). Принцип интерпретации дерева следующий. При определении значения категории первоначально модель обращается к X[4], т.е. пятому в списке категорий (список категорий и номеров, соответствующих им на дереве решений см. в табл. 3). Если значение категории X[4] меньше или равно 0,5, т.е. в действительности принимает значение «0» (при выборке из 144 наблюдений), то стрелка идет влево, если значение категории X[4] больше 0,5, т.е. принимает значение «1», то стрелка идет вправо. В следующих узлах рассматриваются значения категорий X[2] и X[0] и так продолжается до того состояния, пока выбор значения не будет обусловлен только одним (самым близким верхним) узлом. Строка value предоставляет информацию о типе категории: если value принимает значение [n, 0], где $n \geqslant 1$, то в n языках при описанных в верхних узлах условиях порядок VO, если value принимает значение [0, n], где $n \geqslant 1$, то в n языках при описанных в верхних узлах условиях порядок OV.

Таким образом, в присвоении категории порядка прямого объекта и предиката участвуют следующие значения категорий, рассмот-

[1] <http://scikit-learn.org/stable> (дата обращения: 25.05.2015).

Таблица 3

Соответствие категорий и номеров для дерева решений на рис. 4

NDO(VOOV)	**SV**	**VS**	**NDO(SVVS)**	**PrN**	**NPo**	**Inpo**
0	1	2	3	4	5	6
NDO(PrPo)	**NoAdpo**	**GenN**	**NGen**	**NDO(NGen)**	**AdjN**	**NAdj**
7	8	9	10	11	12	13
NDO(NAdj)	**DemN**	**DemSuff**	**NDem**	**DemPref**	**NumN**	**NNum**
14	15	16	17	18	19	20
NDO(NNum)	**RelN**	**NRel**	**DegAdj**	**AdjDeg**	**NDO(DegAdj)**	**InitQP**
21	22	23	24	25	26	27
FinQP	**NoQP**	**AdvSubClause**	**ClauseAdvSub**	**NegV**	**VNeg**	
28	29	30	31	32	33	

ренные в исследовании: PrN — на первом этапе, VS и NDO(VOOV) — на втором этапе, AdvSubClause и ClauseAdvSub — на третьем этапе, NDO(PrPo), NoQP, DemN и GenN — на четвертом этапе, NDO(SVVS) и AdvSubClause — на пятом этапе, NDO(PrPo) и NegV — на шестом этапе, SV — на седьмом этапе, NRel — на восьмом этапе, NDO(NGen) — на девятом и InitQP — на десятом.

Рисунок 4 показывает, что присвоение категорий происходит не по бинарной или тернарной системе — здесь задействовано гораздо больше факторов, что делает импликативные универсалии, состоящие из одного импликанта, неэффективными.

6. Вывод

Результаты кластеризации показывают наличие двух групп порядковых категорий (см. рис. 1) и одной «плавающей» категории. Первая группа включает следующие категории порядка: числительное + имя, прилагательное + имя, демонстратив + имя, модификатор степени + прилагательное, положение вопросительной частицы, отрицание + глагол. Вторая группа состоит из следующих категорий: предикат + прямое дополнение, адлог + имя, посессор + посессум, адвербиальный субординатор + клауза, релятивная клауза + вершина. «Плавающей», т.е. независимой, является категория порядка субъекта и предиката.

Кластеризация по значениям категорий (см. рис. 2, 3) продемонстрировала дробление внутри первой группы на подгруппу с категориями положения отрицания и предиката и положения вопросительной частицы и подгруппу прочих категорий (прилагательное, числительное, демонстратив, модификатор степени).

Результаты корреляционного теста подтвердили данные кластерного анализа. Так, коэффициент корреляции выделил две группы вышеописанных категорий, дополнительно выделив внутри пар категорий ядерные, т.е. такие, значение которых не зависит от прочих категорий внутри группы (предикат + прямой объект и адлог + имя, прилагательное + имя и демонстратив + имя).

Источники

Abdi H., Valentin D. Multiple Correspondence Analysis // N.J. Salkind (ed.). Encyclopedia of Measurement and Statistics. Thousand Oaks (CA): Sage, 2007. P. 651–657.

Baba K., Shibata R., Sibuya M. Partial Correlation and Conditional Correlation as Measures of Conditional Independence // Australian and New Zealand Journal of Statistics. 2004. Vol. 46. No. 4. P. 657–664.

Bisang W. Typology 2. 7th Summer School of the German Linguistic Society. Potsdam, 2002*a*. <http://www.phil-fak.uni-duesseldorf.de/summerschool2002/Bisang2.PDF> (дата обращения: 28.04.2015).

Bisang W. Typology 4. 7th Summer School of the German Linguistic Society. Potsdam, 2002*b*. <http://www.phil-fak.uni-duesseldorf.de/summerschool2002/Bisang4.PDF> (дата обращения: 05.04.2015).

Comrie B. Language Universals and Linguistic Typology. 2nd ed. Chicago: University of Chicago, 1989.

Cramer H. Mathematical Methods of Statistics. Princeton: Princeton University Press, 1946. P. 282.

Croft W. Modern Syntactic Typology. Approaches to Language Typology: Past and Present / M. Shibatani, Th. Bynon (eds). Oxford: Oxford University Press, 1995. P. 85–143.

Dryer M. Coos Word Order. Western Conference on Linguistics. Eugene: University of Oregon, 1983.

Dryer M.S. Large Linguistic Areas and Language Sampling // Studies in Language. 1989. Vol. 13. P. 257–292.

Dryer M.S. The Greenbergian Word Order Correlations // Language. 1992. Vol. 68. P. 81–138.

Dryer M.S. Why Statistical Universals Are Better Than Absolute Universals // Chicago Linguistic Society. 1998. Vol. 33: The Panels. P. 123–145.

Dryer M.S. Word Order // Clause Structure, Language Typology and Syntactic Description. Vol. 1 / T. Shopen (ed.). 2nd ed. Cambridge University Press, 2007.

Dryer M.S., Haspelmath M. (eds). The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology, 2013. <http://wals.info> (дата обращения: 25.02.2015).

Fisher R.A. The Distribution of the Partial Correlation Coefficient // Metron. 1924. Vol. 3. No. 3–4. P. 329–332.

Greenberg J. Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements // Universals of Language. Cambridge, MA: MIT Press, 1963. P. 73–113.

Greenberg J.H. (ed.). Universals of language. 2nd ed. Cambridge, MA: MIT Press, 1966.

Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer-Verlag, 2009. Ch. 15: Random Forests.

Hawkins J.A. Word Order Universals. N.Y.: Academic Press, 1983.

Lucy J. Grammatical Categories and Cognition: A Case Study of the Linguistic Relativity Hypothesis. Cambridge: Cambridge University Press, 1992.

Maddieson I. Patterns of Sounds. Cambridge: Cambridge University Press, 1984.

Mallison G., Blacke J.B. Language Typology. N.Y.: North Holland, 1981.

McCawley J. English As a VSO Language. 1970. Lg. 46.286-99. Reprinted in McCawley 1973e:211-28 // P. Seuren (ed.). Semantic Syntax. Oxford: Clarendon, 1974. P. 74–95.

Meyer D., Zeileis A., Hornik K. VCD: Visualizing Categorical Data. R package version 1.3-2. 2014.

Perkins R.D. Statictical Techniques for Determining Language Sample Size // Studies in Language. 1989. Vol. 13. P. 293–315.

Tomlin S.R. Basic Word Order: Functional Principles. N.Y.: Routledge Library Editions: Linguistics, 1986.