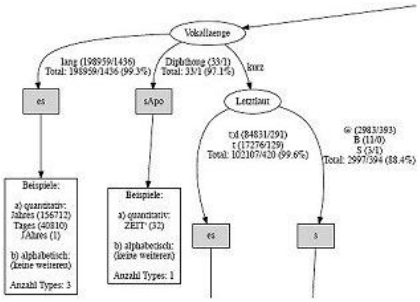
Erschienen auf: grammis 2.0. unter: http://hypermedia.ids-mannheim.de/call/public/korpus.ansicht?v_id=5032
(abgerufen am: 27.06.2016/16:30)

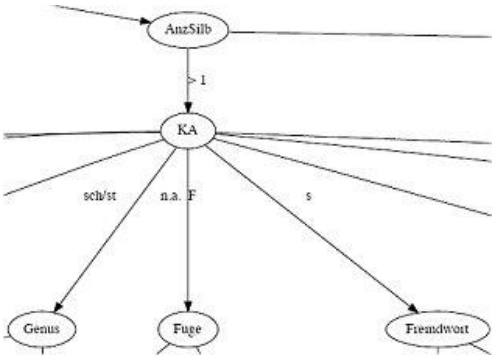
Entscheidungsbäume zur Wahl der Genitivmarkierung

Durch die Modellierung eines Entscheidungsbaumes ist die explorative Vorhersage und das Aufdecken von Regeln für das Verhalten einer bestimmten Variable (hier: der Genitivmarkierung) in Abhängigkeit von verschiedenen Faktoren möglich. Die Klassifikation erfolgt vom Wurzelknoten abwärts über innere Knoten, bis man ein Blatt erreicht, welches die Zielinformation, d.h. eine Ausprägung der im Untersuchungsfokus stehenden Variable, enthält. Jeder Knoten repräsentiert ein Attribut bzw. einen möglichen Einflussfaktor. Je nach Ausprägung des Einflussfaktors folgt man einem unterschiedlichen Zweig, bis man zu einem Blatt gelangt. Die einzelnen Blätter enthalten die jeweiligen Klassifikationen der Variable, die es zu erklären gilt. In unserem Fall bestehen diese aus den Genitivmarkierungsvarianten.

Die GenitivDB präsentiert das optimierte Ergebnis der bereits sechsten Extraktion der Genitivnomina aus dem Deutschen Referenzkorpus (DeReKo-2011-I). Nach jeder Extraktion wurden mehrere Entscheidungsbäume trainiert, die sich zum einen durch statistische Einstellungen und zum anderen durch die Auswahl möglicher Einflussfaktoren unterscheiden. Bei den Modifikationen ging es darum, einerseits die Vorhersagewerte zu steigern bzw. ihr hohes Niveau aufrechtzuerhalten und andererseits den Baum nicht allzu komplex ausfallen zu lassen bzw. seine linguistische Interpretierbarkeit zu sichern. Die Visualisierungen des jeweils letzten Entscheidungsbaumes, der nach der fünften bzw. sechsten Extraktion modelliert wurde, werden hier exemplarisch präsentiert.



Entscheidungsbaum aus Extraktion 5 (Teilansicht; Anklicken öffnet Gesamtbaum))



Entscheidungsbaum aus Extraktion 6 (Teilansicht; Anklicken öffnet Gesamtbaum)

Entscheidungsbäume werden mit Verfahren des maschinellen Lernens auf der Basis von Trainingsdaten erstellt. Um die Wahl der Genitivmarkierungen vorherzusagen, wurde der Algorithmus C4.5, der in der Software WEKA implementiert ist, angewendet. Der Algorithmus testet jeden Faktor daraufhin, ob er die Datenmenge in Gruppen aufteilt, die in sich so wenig Varianz wie möglich aufweisen. Der Faktor, der die Varianz in einer Gruppe am besten erklärt, wird ausgewählt und der Trainingsdatensatz in Teilmengen aufgeteilt. Das Maß für die Aufteilung ist die Kullback-Leibler-Divergenz, die auf der Berechnung der relativen Entropie basiert. Für jede weitere Teilmenge werden die Faktoren mit Hilfe des Maßes bewertet und je nach Bewertung der Faktoren in weitere Teilmengen aufgeteilt. Dieser Prozess wird wiederholt, bis eine Teilmenge keine Varianz mehr aufweist, d.h. lediglich nur noch Fälle einer Klasse enthält oder eine vorgegebene minimale Anzahl von Fällen pro Blatt erreicht ist.

[Sandra Hansen-Morath, Marek Konopka]