

EDAN20

Language Technology

<http://cs.lth.se/edan20/>

Chapter 4: Topics in Information Theory and Machine Learning

Pierre Nugues

Lund University

Pierre.Nugues@cs.lth.se

http://cs.lth.se/pierre_nugues/

September 7, 2016



Why Machine Learning: Early Artificial Intelligence

Early artificial intelligence techniques used introspection to codify knowledge, often in the form of rules.

Expert systems, one of the most notable applications of traditional AI, were entirely based on the competence of experts.

This has two major drawbacks:

- Need of an expertise to understand and explain the rules
- Bias introduced by the expert



Why Machine Learning: What has Changed

Now terabytes of data available.

Makes it impossible to understand such volumes of data, organize them using manually-crafted rules.

Triggered a major move to empirical and statistical techniques.

In fact, most machine-learning techniques come from traditional statistics.

Applications in natural language processing, medicine, banking, online shopping, image recognition, etc.

The success of companies like Google, Facebook, Amazon, and Netflix, not to mention Wall Street firms and industries from manufacturing and retail to healthcare, is increasingly driven by better tools for extracting meaning from very large quantities of data. 'Data Scientist' is now the hottest job title in Silicon Valley.

– Tim O'Reilly



Some Definitions

- 1 Machine learning always starts with **data sets**: a collection of objects or observations.
- 2 Machine-learning algorithms can be classified along two main lines: **supervised** and **unsupervised** classification.
- 3 Supervised algorithms need a **training set**, where the objects are described in terms of attributes and belong to a known class or have a known output.
- 4 The performance of the resulting classifier is measured against a **test set**.
- 5 We can also use N -fold cross validation, where the test set is selected randomly from the training set N times, usually 10.
- 6 Unsupervised algorithms consider objects, where no class is provided.
- 7 Unsupervised algorithms learn regularities in data sets.



A Data Set: *Salammbô*

A corpus is a collection – a body – of texts.

French original

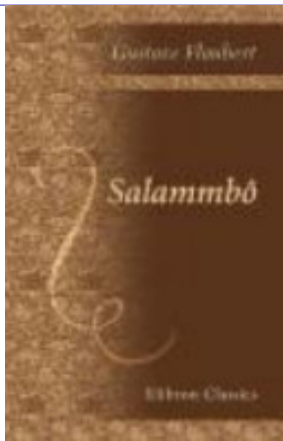
English translation

GUSTAVE FLAUBERT

SALAMMBÔ

ÉDITION DÉFINITIVE
AVEC DES DOCUMENTS NOUVEAUX

PARIS
G. CHARPENTIER, ÉDITEUR
13, RUE DE CHEVREUIL-SAINTE-GENEVIEVE, 13
—
1883
Tous droits réservés



Supervised Learning

Letter counts from *Salammbô*

| Chapter | French | | English | |
|------------|--------------|-------|--------------|-------|
| | # characters | # A | # characters | # A |
| Chapter 1 | 36,961 | 2,503 | 35,680 | 2,217 |
| Chapter 2 | 43,621 | 2,992 | 42,514 | 2,761 |
| Chapter 3 | 15,694 | 1,042 | 15,162 | 990 |
| Chapter 4 | 36,231 | 2,487 | 35,298 | 2,274 |
| Chapter 5 | 29,945 | 2,014 | 29,800 | 1,865 |
| Chapter 6 | 40,588 | 2,805 | 40,255 | 2,606 |
| Chapter 7 | 75,255 | 5,062 | 74,532 | 4,805 |
| Chapter 8 | 37,709 | 2,643 | 37,464 | 2,396 |
| Chapter 9 | 30,899 | 2,126 | 31,030 | 1,993 |
| Chapter 10 | 25,486 | 1,784 | 24,843 | 1,627 |
| Chapter 11 | 37,497 | 2,641 | 36,172 | 2,375 |
| Chapter 12 | 40,398 | 2,766 | 39,552 | 2,560 |
| Chapter 13 | 74,105 | 5,047 | 72,545 | 4,597 |
| Chapter 14 | 76,725 | 5,312 | 75,352 | 4,871 |
| Chapter 15 | 18,317 | 1,215 | 18,031 | 1,119 |



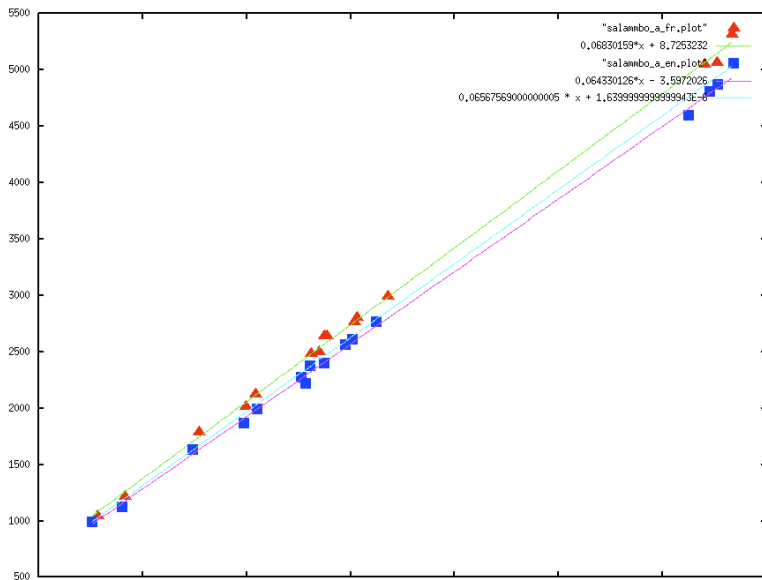
Classification Data Set

A binary classification.

| | # char. | # A | class (y) | # char. | # A | class (y) |
|------------|---------|-------|-----------|---------|-------|-----------|
| Chapter 1 | 36,961 | 2,503 | 1 | 35,680 | 2,217 | 0 |
| Chapter 2 | 43,621 | 2,992 | 1 | 42,514 | 2,761 | 0 |
| Chapter 3 | 15,694 | 1,042 | 1 | 15,162 | 990 | 0 |
| Chapter 4 | 36,231 | 2,487 | 1 | 35,298 | 2,274 | 0 |
| Chapter 5 | 29,945 | 2,014 | 1 | 29,800 | 1,865 | 0 |
| Chapter 6 | 40,588 | 2,805 | 1 | 40,255 | 2,606 | 0 |
| Chapter 7 | 75,255 | 5,062 | 1 | 74,532 | 4,805 | 0 |
| Chapter 8 | 37,709 | 2,643 | 1 | 37,464 | 2,396 | 0 |
| Chapter 9 | 30,899 | 2,126 | 1 | 31,030 | 1,993 | 0 |
| Chapter 10 | 25,486 | 1,784 | 1 | 24,843 | 1,627 | 0 |
| Chapter 11 | 37,497 | 2,641 | 1 | 36,172 | 2,375 | 0 |
| Chapter 12 | 40,398 | 2,766 | 1 | 39,552 | 2,560 | 0 |
| Chapter 13 | 74,105 | 5,047 | 1 | 72,545 | 4,597 | 0 |
| Chapter 14 | 76,725 | 5,312 | 1 | 75,352 | 4,871 | 0 |
| Chapter 15 | 18,317 | 1,215 | 1 | 18,031 | 1,119 | 0 |



Separating Classes



Supervised Learning: Fisher's Iris data set (1936)

180 MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS

Table I

| <i>Iris setosa</i> | | | | <i>Iris versicolor</i> | | | | <i>Iris virginica</i> | | | |
|--------------------|-------------|--------------|-------------|------------------------|-------------|--------------|-------------|-----------------------|-------------|--------------|-------------|
| Sepal length | Sepal width | Petal length | Petal width | Sepal length | Sepal width | Petal length | Petal width | Sepal length | Sepal width | Petal length | Petal width |
| 5.1 | 3.5 | 1.4 | 0.2 | 7.0 | 3.2 | 4.7 | 1.4 | 6.3 | 3.3 | 6.0 | 2.5 |
| 4.9 | 3.0 | 1.4 | 0.2 | 6.4 | 3.2 | 4.5 | 1.5 | 5.8 | 2.7 | 5.1 | 1.9 |
| 4.7 | 3.2 | 1.3 | 0.2 | 6.9 | 3.1 | 4.9 | 1.5 | 7.1 | 3.0 | 5.9 | 2.1 |
| 4.6 | 3.1 | 1.5 | 0.2 | 5.5 | 2.3 | 4.0 | 1.3 | 6.3 | 2.9 | 5.6 | 1.8 |
| 5.0 | 3.6 | 1.4 | 0.2 | 6.5 | 2.8 | 4.6 | 1.5 | 6.5 | 3.0 | 5.8 | 2.2 |
| 5.4 | 3.9 | 1.7 | 0.4 | 5.7 | 2.8 | 4.5 | 1.3 | 7.6 | 3.0 | 6.6 | 2.1 |
| 4.6 | 3.4 | 1.4 | 0.3 | 6.3 | 3.3 | 4.7 | 1.6 | 4.9 | 2.5 | 4.5 | 1.7 |
| 5.0 | 3.4 | 1.5 | 0.2 | 4.9 | 2.4 | 3.3 | 1.0 | 7.3 | 2.9 | 6.3 | 1.8 |
| 4.4 | 2.9 | 1.4 | 0.2 | 6.6 | 2.9 | 4.6 | 1.3 | 6.7 | 2.5 | 5.8 | 1.8 |
| 4.9 | 3.1 | 1.5 | 0.1 | 5.2 | 2.7 | 3.9 | 1.4 | 7.2 | 3.6 | 6.1 | 2.5 |
| 5.4 | 3.7 | 1.5 | 0.2 | 5.0 | 2.0 | 3.5 | 1.0 | 6.5 | 3.2 | 5.1 | 2.0 |
| 4.8 | 3.4 | 1.6 | 0.2 | 5.9 | 3.0 | 4.2 | 1.5 | 6.4 | 2.7 | 5.3 | 1.9 |
| 4.8 | 3.0 | 1.4 | 0.1 | 6.0 | 2.2 | 4.0 | 1.0 | 6.8 | 3.0 | 5.5 | 2.1 |
| 4.3 | 3.0 | 1.1 | 0.1 | 6.1 | 2.9 | 4.7 | 1.4 | 5.7 | 2.5 | 5.0 | 2.0 |
| 5.8 | 4.0 | 1.2 | 0.2 | 5.6 | 2.9 | 3.6 | 1.3 | 5.8 | 2.8 | 5.1 | 2.4 |
| 5.7 | 4.4 | 1.5 | 0.4 | 6.7 | 3.1 | 4.4 | 1.4 | 6.4 | 3.2 | 5.3 | 2.3 |
| 5.4 | 3.9 | 1.3 | 0.4 | 5.6 | 3.0 | 4.5 | 1.5 | 6.5 | 3.0 | 5.5 | 1.8 |
| 5.1 | 3.5 | 1.4 | 0.3 | 5.8 | 2.7 | 4.1 | 1.0 | 7.7 | 3.8 | 6.7 | 2.2 |
| 5.7 | 3.8 | 1.7 | 0.3 | 6.2 | 2.2 | 4.5 | 1.5 | 7.7 | 2.6 | 6.9 | 2.3 |
| 5.1 | 3.8 | 1.5 | 0.3 | 5.6 | 2.5 | 3.9 | 1.1 | 6.0 | 2.2 | 5.0 | 1.5 |
| 5.4 | 3.4 | 1.7 | 0.2 | 5.9 | 3.2 | 4.8 | 1.8 | 6.9 | 3.2 | 5.7 | 2.3 |
| 5.1 | 3.7 | 1.5 | 0.4 | 6.1 | 2.8 | 4.0 | 1.3 | 5.6 | 2.8 | 4.9 | 2.0 |

Berkson's Data Set (1944)

| Drug concentration | Number exposed | Survive Class 0 | Die Class 1 | Mortality rate | Expected mortality |
|--------------------|----------------|-----------------|-------------|----------------|--------------------|
| 40 | 462 | 352 | 110 | .2359 | .2206 |
| 60 | 500 | 301 | 199 | .3980 | .4339 |
| 80 | 467 | 169 | 298 | .6380 | .6085 |
| 100 | 515 | 145 | 370 | .7184 | .7291 |
| 120 | 561 | 102 | 459 | .8182 | .8081 |
| 140 | 469 | 69 | 400 | .8529 | .8601 |
| 160 | 550 | 55 | 495 | .9000 | .8952 |
| 180 | 542 | 43 | 499 | .9207 | .9195 |
| 200 | 479 | 29 | 450 | .9395 | .9366 |
| 250 | 497 | 21 | 476 | .9577 | .9624 |
| 300 | 453 | 11 | 442 | .9757 | .9756 |

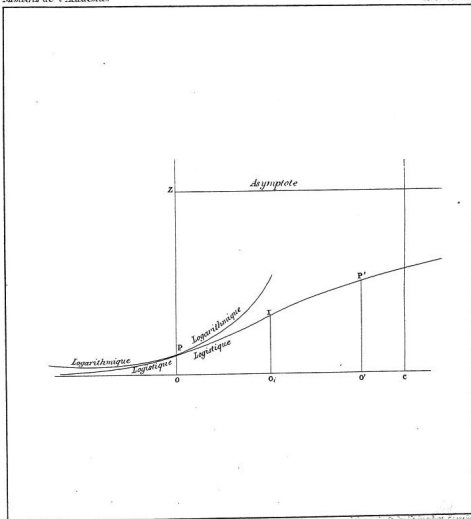
Table: A data set. Adapted and simplified from the original article that described how to apply logistic regression to classification by Joseph Berkson, Application of the Logistic Function to Bio-Assay. *Journal of the American Statistical Association* (1944).



Another Model, the Logistic Curve (Verhulst)

Mémoires de l'Académie.

Tome XVIII.



Mémoire sur la population par M. P. Verhulst.

$$\text{Logistic}(x) = \frac{1}{1 + e^{-x}}$$

$$\begin{aligned} \hat{y}(x) &= \text{Logistic}(w \cdot x) \\ &= \frac{1}{1 + e^{-w \cdot x}} \end{aligned}$$

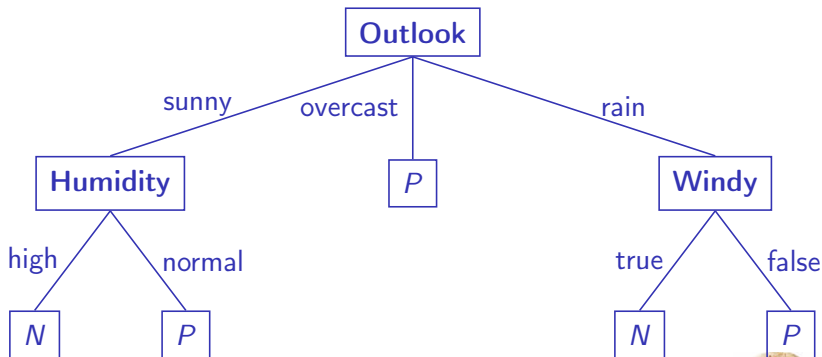


Objects, Attributes, and Classes. After Quinlan (1986)

| Object | Attributes | | | | Class |
|--------|------------|-------------|----------|-------|----------|
| | Outlook | Temperature | Humidity | Windy | |
| 1 | Sunny | Hot | High | False | <i>N</i> |
| 2 | Sunny | Hot | High | True | <i>N</i> |
| 3 | Overcast | Hot | High | False | <i>P</i> |
| 4 | Rain | Mild | High | False | <i>P</i> |
| 5 | Rain | Cool | Normal | False | <i>P</i> |
| 6 | Rain | Cool | Normal | True | <i>N</i> |
| 7 | Overcast | Cool | Normal | True | <i>P</i> |
| 8 | Sunny | Mild | High | False | <i>N</i> |
| 9 | Sunny | Cool | Normal | False | <i>P</i> |
| 10 | Rain | Mild | Normal | False | <i>P</i> |
| 11 | Sunny | Mild | Normal | True | <i>P</i> |
| 12 | Overcast | Mild | High | True | <i>P</i> |
| 13 | Overcast | Hot | Normal | False | <i>P</i> |
| 14 | Rain | Mild | High | True | <i>N</i> |



Classifying Objects with Decision Trees. After Quinlan (1986)



Supervised Machine-Learning Algorithms

Linear classifiers:

- 1 Perceptron
- 2 Support vector machines
- 3 Logistic regression



Classification

We represent classification using a threshold function (a variant of the signum function):

$$H(\mathbf{w} \cdot \mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{w} \cdot \mathbf{x} \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

The classification function associates P with 1 and N with 0.
We want to find the separating hyperplane:

$$\begin{aligned} \hat{y}(\mathbf{x}) &= H(\mathbf{w} \cdot \mathbf{x}) \\ &= H(w_0 + w_1x_1 + w_2x_2 + \dots + w_nx_n), \end{aligned}$$

given a data set of q examples: $DS = \{(1, x_1^j, x_2^j, \dots, x_n^j, y^j) | j : 1..q\}$.

We use $x_0 = 1$ to simplify the equations.

For a binary classifier, y has then two possible values $\{0, 1\}$ corresponding in our example to $\{\text{French, English}\}$.



Notation

- A feature vector (predictors): \mathbf{x} , and feature matrix: \mathbf{X} ;
- The class: y and the class vector: \mathbf{y} ;
- The predicted class (response): \hat{y} , and predicted class vector: $\hat{\mathbf{y}}$

$$\mathbf{X} = \begin{bmatrix} \text{Sunny} & \text{Hot} & \text{High} & \text{False} \\ \text{Sunny} & \text{Hot} & \text{High} & \text{True} \\ \text{Overcast} & \text{Hot} & \text{High} & \text{False} \\ \text{Rain} & \text{Mild} & \text{High} & \text{False} \\ \text{Rain} & \text{Cool} & \text{Normal} & \text{False} \\ \text{Rain} & \text{Cool} & \text{Normal} & \text{True} \\ \text{Overcast} & \text{Cool} & \text{Normal} & \text{True} \\ \text{Sunny} & \text{Mild} & \text{High} & \text{False} \\ \text{Sunny} & \text{Cool} & \text{Normal} & \text{False} \\ \text{Rain} & \text{Mild} & \text{Normal} & \text{False} \\ \text{Sunny} & \text{Mild} & \text{Normal} & \text{True} \\ \text{Overcast} & \text{Mild} & \text{High} & \text{True} \\ \text{Overcast} & \text{Hot} & \text{Normal} & \text{False} \\ \text{Rain} & \text{Mild} & \text{High} & \text{True} \end{bmatrix}; \mathbf{y} = \begin{bmatrix} \text{N} \\ \text{N} \\ \text{P} \\ \text{P} \\ \text{P} \\ \text{N} \\ \text{P} \\ \text{N} \\ \text{P} \\ \text{P} \\ \text{P} \\ \text{P} \\ \text{P} \\ \text{N} \end{bmatrix}$$



Converting Symbolic Attributes into Numerical Vectors

Linear classifiers are numerical systems.

Symbolic – nominal – attributes are mapped onto vectors of binary values.
A conversion of the weather data set.

| Object | Attributes | | | | | | | | | | Class |
|--------|------------|----------|------|-------------|------|------|----------|--------|-------|-------|-------|
| | Outlook | | | Temperature | | | Humidity | | Windy | | |
| | Sunny | Overcast | Rain | Hot | Mild | Cool | High | Normal | True | False | |
| 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | N |
| 2 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | N |
| 3 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | P |
| 4 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | P |
| 5 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | P |
| 6 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | N |
| 7 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | P |
| 8 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | N |
| 9 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | P |
| 10 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | P |
| 11 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | P |
| 12 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | P |
| 13 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | P |
| 14 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | N |



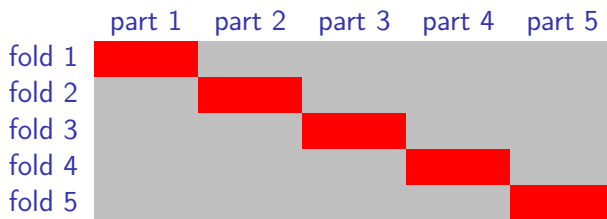
Evaluation

- The standard evaluation procedure is to train the classifier on a training set and evaluate the performance on a test set.
- When we have only one set, we divide it in two subsets: the training set and the test set (or holdout data).
- The split can be 90–10 or 80–20
- This often optimizes the classifier for a specific test set and creates an overfit



Cross Validation

- A N -fold cross validation mitigates the overfit
- The set is partitioned into N subsets, $N = 5$ for example, one of them being the test set (red) and the rest the training set (gray).
- The process is repeated N times with a different test set: N folds



At the extreme, leave-one-out cross-validation



Model Selection

- Validation can apply to one classification method
- We can use it to select a classification method and its parametrization.
- Needs three sets: training set, development set, and test set.



Evaluation

There are different kinds of measures to evaluate the performance of machine learning techniques, for instance:

- Precision and recall in information retrieval and natural language processing;
- The *receiver operating characteristic* (ROC) in medicine.

| | Positive examples: P | Negative examples: N |
|-------------------|------------------------|------------------------|
| Classified as P | True positives: A | False positives: B |
| Classified as N | False negatives: C | True negatives: D |

More on the receiver operating characteristic here: [http:](http://en.wikipedia.org/wiki/Receiver_operating_characteristic)

[//en.wikipedia.org/wiki/Receiver_operating_characteristic](http://en.wikipedia.org/wiki/Receiver_operating_characteristic)



Recall, Precision, and the F-Measure

The **accuracy** is $\frac{|A \cup D|}{|P \cup N|}$.

Recall measures how much relevant examples the system has classified correctly, for P :

$$\text{Recall} = \frac{|A|}{|A \cup C|}.$$

Precision is the accuracy of what has been returned, for P :

$$\text{Precision} = \frac{|A|}{|A \cup B|}.$$

Recall and precision are combined into the **F-measure**, which is defined as the harmonic mean of both numbers:

$$F = \frac{2 \cdot \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$



Measuring Quality: The Confusion Matrix

A task in natural language processing: Identify the parts of speech (POS) of words.

Example: *The can rusted*

- The human: *The*/art (DT) *can*/noun (NN) *rusted*/verb (VBD)
- The POS tagger: *The*/art (DT) *can*/modal (MD) *rusted*/verb (VBD)

| ↓Correct | Tagger → | | | | | | | | | |
|----------|----------|------|------|------|------|------|------|------|------|------|
| | DT | IN | JJ | NN | RB | RP | VB | VBD | VBG | VBN |
| DT | 99.4 | 0.3 | – | – | 0.3 | – | – | – | – | – |
| IN | 0.4 | 97.5 | – | – | 1.5 | 0.5 | – | – | – | – |
| JJ | – | 0.1 | 93.9 | 1.8 | 0.9 | – | 0.1 | 0.1 | 0.4 | 1.5 |
| NN | – | – | 2.2 | 95.5 | – | – | 0.2 | – | 0.4 | – |
| RB | 0.2 | 2.4 | 2.2 | 0.6 | 93.2 | 1.2 | – | – | – | – |
| RP | – | 24.7 | – | 1.1 | 12.6 | 61.5 | – | – | – | – |
| VB | – | – | 0.3 | 1.4 | – | – | 96.0 | – | – | 0.2 |
| VBD | – | – | 0.3 | – | – | – | – | 94.6 | – | 4.8 |
| VBG | – | – | 2.5 | 4.4 | – | – | – | – | 93.0 | – |
| VBN | – | – | 4.6 | – | – | – | – | 4.3 | – | 90.6 |

After Franz (1996, p. 124)