Automatic animacy classification
for Dutch

Jelke Bloem, Gosse Bouma

Noun animacy

- Sentience of the referent

sister – participant – carpenter – dude – northener

cat – angel – dragon – bacteria

oak – robot – community – government

fountain – second – observation – music – fog

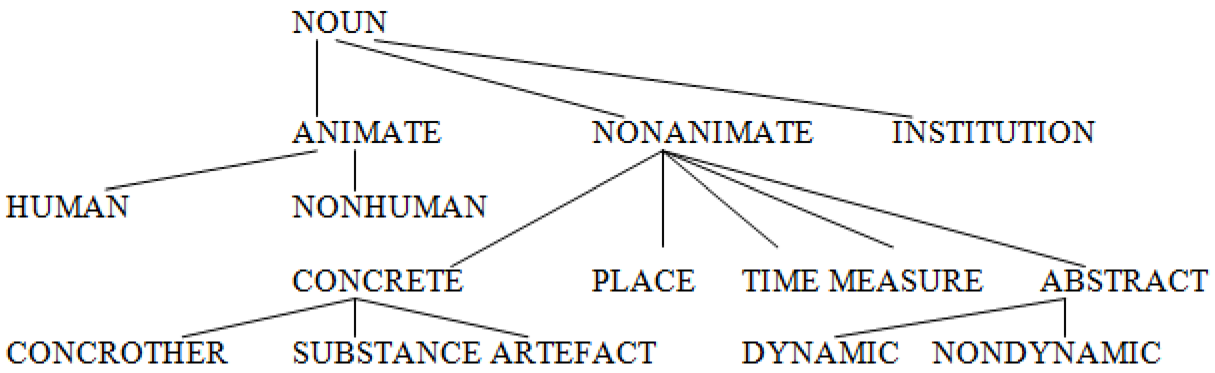
- Animacy hierarchy

Animacy hierarchy

- Humans (OConnor et al., 2004)
- Other animates

  ORGANIZATIONS, ANIMALS, INTELLIGENT MACHINES, VEHICLES.
- Inanimates

  CONCRETE INANIMATE, NON-CONCRETE INANIMATE, PLACE, TIME



NOUN

ANIMATE NONANIMATE INSTITUTION

HUMAN NONHUMAN

CONCRETE PLACE TIME MEASURE ABSTRACT

CONCROTHER SUBSTANCE ARTEFACT DYNAMIC NONDYNAMIC

(Martin et al., 2005)

- HUMAN > NONHUMAN > NONANIMATE

Animacy and grammaticality

- The spoon **which** is on the table is mine.
- * The man **which** is sitting on the table is my friend.

- * The spoon **who** is on the table is mine.
- The man **who** is sitting on the table is my friend.

who refers to ANIMATE, **which** refers to INANIMATE

- Cut-off point

Animacy and sentence processing

- Dative alternation (Bresnan et al. 2007)

She gave a push to the car . (prepositional dative)

She gave the car a push. (double object)

She gave a toy to the child. (prepositional dative)

She gave the child a toy. (double object)

- For inanimate recipients, the double object construction is used more often

Automatic animacy classification goals

- Corpus annotation
- Use in language technology
  - e.g. Automatic translation:
    *De man **die** op de tafel zat*
    die = that, which, **who**, those, these
    The man **who** sat on the table
  - Anaphora resolution (Orasan and Evans, 2007)
    The tree fell on the man. He survived.
  - Better parsing (Øvrelid, 2009)

Animacy in natural language processing

› Few animacy resources are available (Zaenen, 2004)

› Therefore, few tools make use of animacy

› A few animacy classifiers were made, none for Dutch

› Dutch resources:

Cornetto (lexical-semantic database)
Lassy Large (automatically annotated corpus), 1.5 billion words

Animacy classification task

› For any noun, decide whether it refers to a human, nonhuman animate or inanimate entity

› Classification features
  - World knowledge?
  - Morphology?
  - Context?

Animacy classification task

- World knowledge (Orasan and Evans, 2007)
  - Lexical-semantic database (WordNet)

    poet -> writer -> communicator -> **person**

    wikipedian -> ???

- Morphology (Baker and Brew, 2008)

  诗 人

  poetry **person**

  or: Case marking

Context features

(Øvrelid, 2009)

- Animates prefer the agent role & subject position
- Inanimates prefer the patient role & object position
- Genitive case
    - das Haus mein**es** Vater**s**
- Reflexive
    - **The teacher** hurt **himself**

Context features

- Lexical association features: Verbs

  The doctor **thought** John was right.

  The banana **thought** John was right.

- Adjectives

  The **lazy** thief.

  The **lazy** hurricane.

Animacy: Data

- Word lemmas and their animacy from Cornetto
- Verb-argument relations from Lassy Large corpus

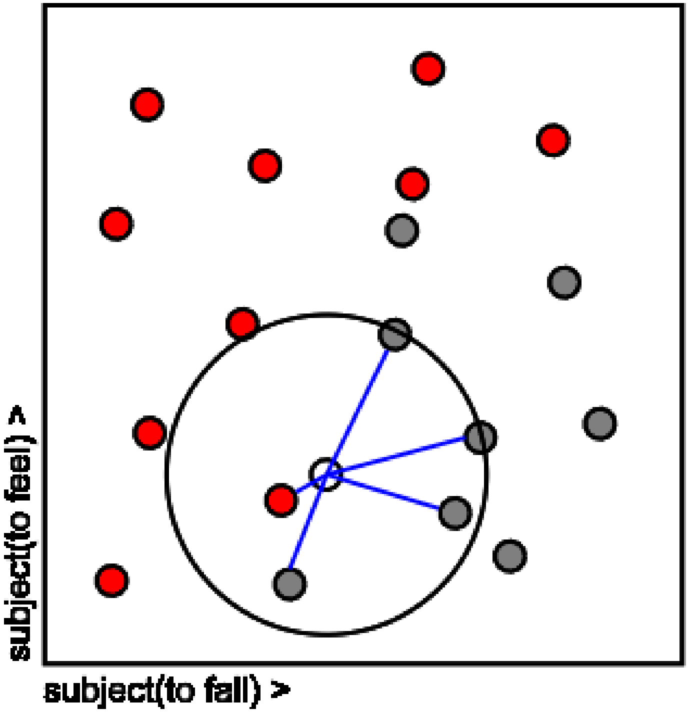
```
<noun animacy="nonanimate">gevoel</noun>
<noun animacy="nonanimate">IJsselmeer</noun>
<noun animacy="nonanimate">noord</noun>
<noun animacy="nonanimate">paasei</noun>
<noun animacy="human">doctor</noun>
<noun animacy="human">Engelsman</noun>
<noun animacy="human">roker</noun>
<noun animacy="human">symfonieorkest</noun>
<noun animacy="nonhuman">fuchsia</noun>
<noun animacy="nonhuman">pony</noun>
<noun animacy="nonhuman">yeti</noun>
```

```
85#blijf|intransitive|su#gevoel
298#ontsta|intransitive|su#gevoel
1#schrijf|transitive|su#gevoel
8#rest|intransitive|su#gevoel

7#ontdek|transitive|su#Engelsman
4#ontwerp|transitive|su#Engelsman
3#overschat|transitive|su#Engelsman
```

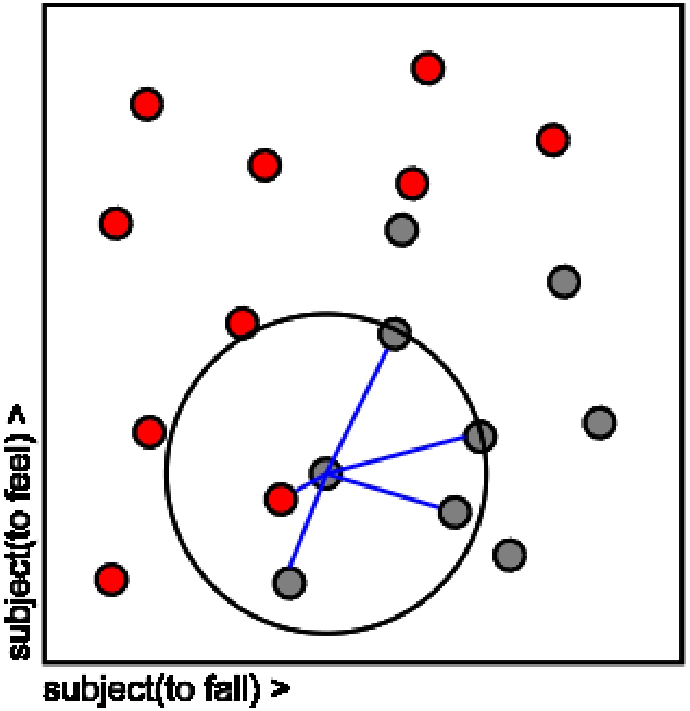
Classification procedure

- K-nearest neighbor (TiMBL)
- Each noun is a feature vector
- Classify new instances based on most similar (nearest) noun in multidimensional feature space



Classification procedure

- K-nearest neighbor (TiMBL)
- Each noun is a feature vector
- Classify new instances based on most similar (nearest) noun in multidimensional feature space



- 4 of 5 neighbors are inanimate
  - The inanimate class is assigned

Feature values: Association strength

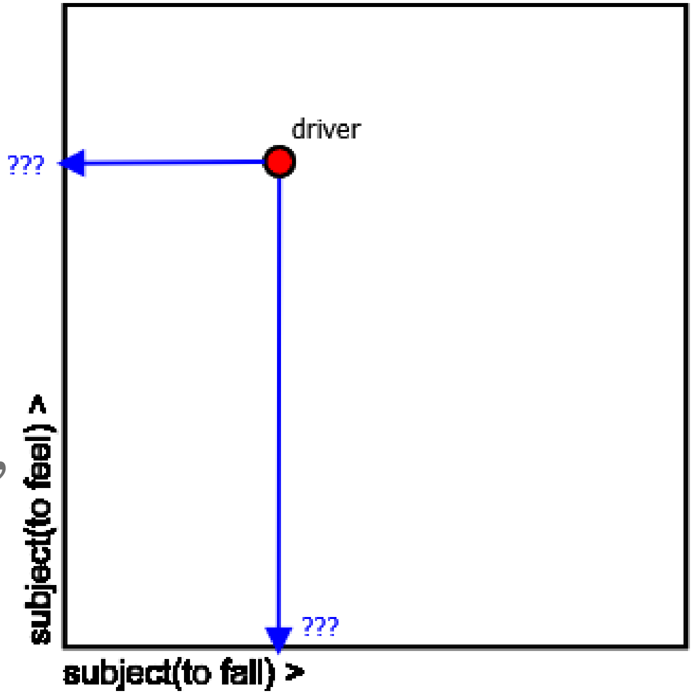
- Noun-verb association

```
Subj(ben): bestuurder 125
Subj(rij): bestuurder 12
```

- Which is more interesting?

- Pointwise Mutual Information, Fisher's Exact Test

Feature-noun pairs that co-occur more often than would be expected by chance



Association strength

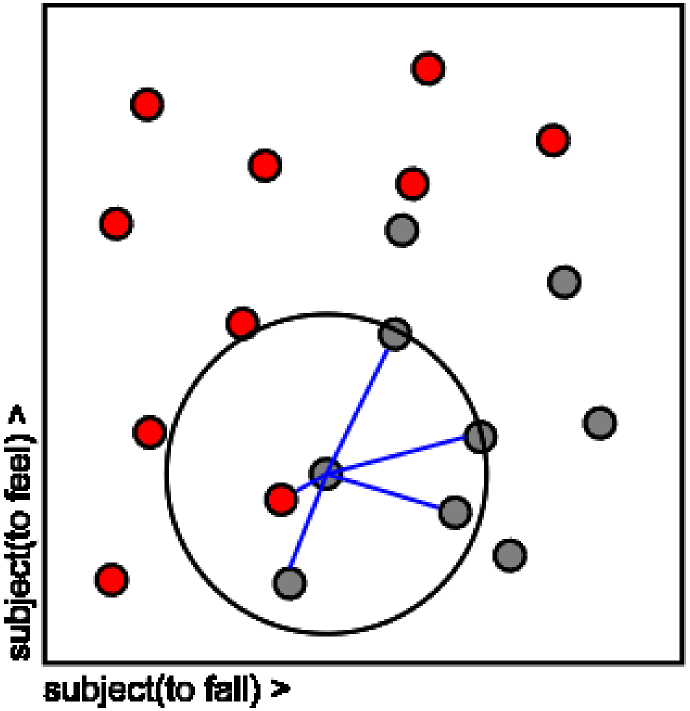
“gevoel” (*feeling*, inanimate) subject relation strength (Fisher’s)

0.000000000000000	ontsta	*arise*
0.000000000000830	heb	*have*
0.000000000002380	speel	*play*
0.000000000501125	ben	*be*
0.000000003404273	zeg	*say*

0.731409478841741	krijg	*get*
0.823487761949459	spreek	*speak*
0.853510038160385	neem	*take*
0.902189553992116	ken	*know*
1.000000000002866	schrijf	*write*

Classification procedure

- K-nearest neighbor (TiMBL)
- Feature values are association scores
- Evaluate by classifying unseen nouns according to these features



Results: Features

Features	Accuracy
Baseline	80.92%
Object/subject ratio	81.09%
Verb subject relations	91.06%
Verb object relations	91.20%
Adjective relations	88.91%
Subj+Obj+Adj	**93.34%**

Baseline: Classify everything as the majority class

Ten-fold cross validation accuracy scores

Noun frequency

- Classifying low-frequency nouns is generally more difficult

Frequency cutoff	**Baseline**	**Accuracy**	**Number of nouns**
>0	76.68%	83.39%	30.950
>1	78.16%	90.27%	16.454
>10	80.92%	**93.34%**	12.168
>100	84.00%	91.22**%**	6.276
>1000	88.99%	88.62%	1.671

Results: Class confusion

- Classification errors

Predicted ->	Human	Nonhuman	Nonanimate
Human	151	0	24
Nonhuman	0	2	53
Nonanimate	1	3	982

- NONHUMAN class is only chosen correctly twice!

Results: Two-way classification

- Human/Nonhuman

Features	Accuracy
Baseline	85.57%
All	**98.03%**

Pred ->	Human	NonH
Human	152	23
NonH	1	1040

- Animate/Inanimate

Features	Accuracy
Baseline	80.92%
All	**92.52%**

Pred ->	Anim.	Inanim.
Anim.	155	75
Inanim.	16	970

Discussion

- Can classify over 90% of Dutch nouns correctly
  - The Cornetto “nonhuman animate” class cannot be classified well
- Corpus creation/annotation
- Applications (parser, anaphora resolution)

- Token-based instead of lemma-based (DutchSemCor)?
- Reduce resource requirements
  - Incorporate morphology, seed set

References

K. Baker and C.Brew. Multilingual animacy classification by sparse logistic regression. *Ohio State Working Papers in Linguistics*, 2008.

J. Bresnan, A. Cueni, T. Nikitina, and R.H. Baayen. Predicting the dative alternation. *Cognitive foundations of interpretation*, pages 69–94, 2007.

W. Martin, I. Maks, S. Bopp, and M. Groot. RBN-documentatie. *Report, TST Centrale*, 2005.

M.C. O'Connor, A. Anttila, V. Fong, and J. Maling. Differential possessor expression in English: Re-evaluating animacy and topicality effects. In *Annual Meeting of the Linguistic Society of America,* January, pages 9–11, 2004.

C. Orasan and R. Evans. NP animacy identification for anaphora resolution. *Journal of Artificial Intelligence Research*, 29(1):79–103, 2007.

L. Øvrelid. Empirical evaluations of animacy annotation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2009.

A. Zaenen, J. Carletta, G. Garretson, J. Bresnan, A. Koontz-Garboden, T. Nikitina, M.C. O'Connor, and T. Wasow. Animacy Encoding in English: why and how. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, pages 118-125. Association for Computational Linguistics, 2004.

Cornetto lexical-semantic database: http://www2.let.vu.nl/oz/cltl/cornetto/

Lassy Large corpus: http://www.let.rug.nl/vannoord/Lassy/

Questions?

Thank you for your attention

Classes

Human

Brabander

Eerste-Kamerlid

afstammeling

begeleidingsteam

drieling

ex-burgemeester

geallieerden

haantje-de-voorste

juf

oermens

racist

tachtiger

Nonhuman

ANWB

appelboom

brandweer

cycloop

dienstensector

embryo

familie

ijsbergsla

maatjesharing

microbe

olifant

snackbar

vrouwenrechten

Inanimate

Groningen

Koninginnedag

appel

belastingkantoor

compassie

friettent

gebarentaal

keel

orkaan

robot

sneltrein

terrorisme

zeewier

Russian case marking

Fraser and Corbett (1995)

pervogo (acc=gen)	*studenta (acc=gen)*
first	student

'the first student'

pervyj (acc=nom)	*zakon (acc=nom)*
first	law

'the first law'

Dutch Wh-clefts

a. ***Wat*** *ik leuk vind, is die tafel(*GEN=COMM,-ANIMATE*)*
what i like, is that table

b. ***Wat*** *ik leuk vind, is dat huis(*GEN=NEUT,-ANIMATE*)*
what i like, is that house

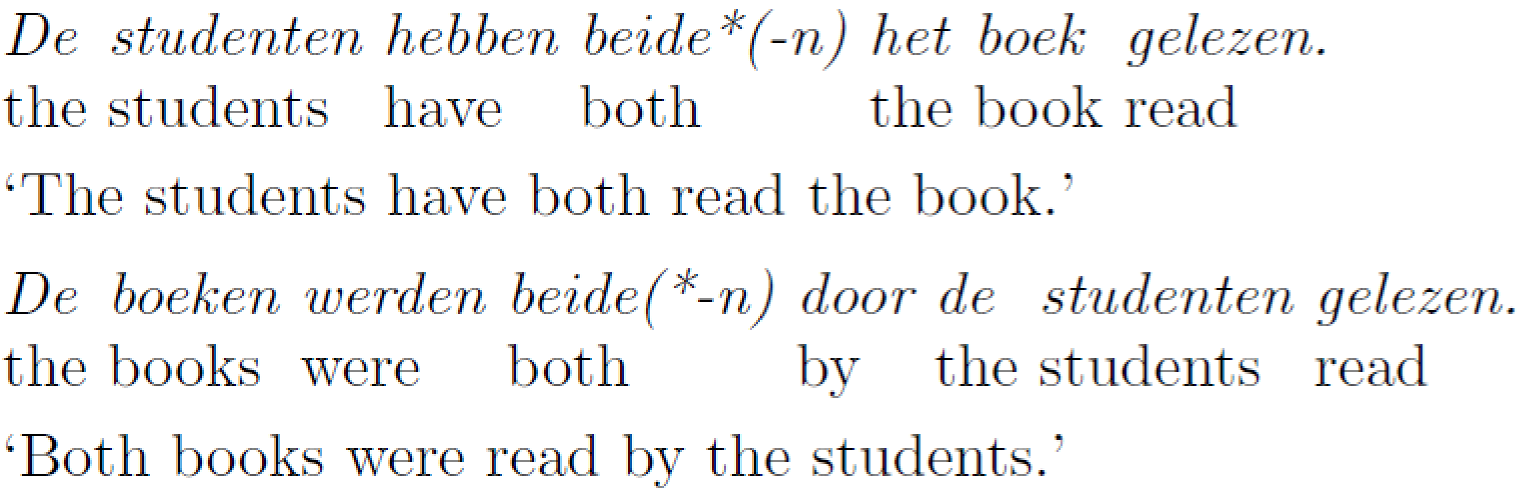
c. ***Wie*** *ik leuk vind, is dat kind(*GEN=NEUT,+ANIMATE*)*
who i like, is that child

d. ***Wie*** *ik leuk vind, is die vrouw(*GEN=COMM,+ANIMATE*)*
who i like, is that woman

› Found no good counter-examples in corpus search

Dutch quantifier suffixes

(de Swart et al., 2008)



De studenten hebben beide(-n) het boek gelezen.*
the students have both the book read
'The students have both read the book.'

De boeken werden beide(-n) door de studenten gelezen.*
the books were both by the students read
'Both books were read by the students.'

- › In written Dutch

Fisher’s Exact Test: Contingency table

- The Fisher’s exact test is calculated using tables
- Totals are fixed

The noun “gevoel” (*feeling*) as a subject of the verb “ontstaan” (*to start, to arise*)

	gevoel	¬gevoel	Row totals
ontstaan	**298**	5927	**6225**
¬ontstaan	405	111952	112357
Column totals	**703**	117879	**118582**

$p < 0.00001$

Dependence and independence

- The p-value can go both ways: Association strength

The noun "gevoel" (*feeling*) as a subject of the verb "schrijven" (*to write*)

	gevoel	¬gevoel	Row totals
schrijven	**1**	299	**300**
¬schrijven	702	117578	118282
Column totals	**703**	117879	**118582**

$p > 0.99999$

Association strength

- This p-value can be used as a measure of association strength
- A low value indicates a strong association, a high value indicates none

- Because the totals are fixed, you cannot compare p-values from samples of different sizes

Fisher’s exact test Hypothesis

- H0: The noun x and the verb y are independent in subject relations
- H1: The noun x occurs as a subject of the verb y more often than would be expected by chance

Calculating the value

- The p-value expresses the total probability of the observed distribution (table) and all the more extreme ones

	gevoel	¬gevoel
ontstaan	**298**	5927
¬ontstaan	405	111952

	gevoel	¬gevoel
ontstaan	**299**	5926
¬ontstaan	404	111951

	gevoel	¬gevoel
ontstaan	**300**	5925
¬ontstaan	403	111950

	gevoel	¬gevoel
ontstaan	**301**	5924
¬ontstaan	402	111949

Calculating the value

	gevoel	¬gevoel	totals
ontstaan	298	5927	6225
¬ontstaan	405	111952	112357
totals	703	117879	118582

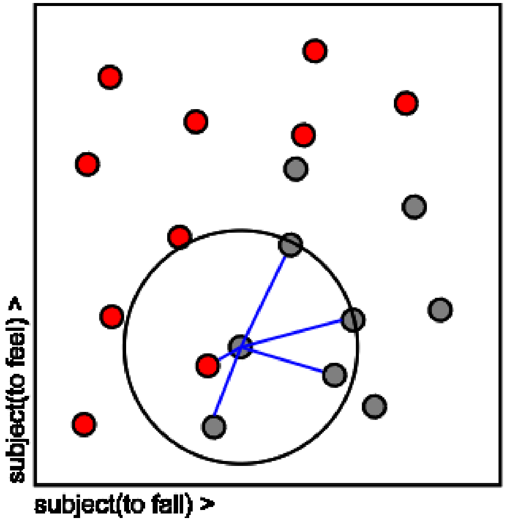
- $P(n) = \dfrac{6225! \ast 112357! \ast 703! \ast 117879!}{298! \ast 5927! \ast 405! \ast 111952! \ast 118582!}$
- $P(n+1) = \dfrac{6225! \ast 112357! \ast 703! \ast 117879!}{299! \ast 5926! \ast 404! \ast 111951! \ast 118582!}$
- etc
- p = $P(n) + P(n+1) + P(n+2) + \ \ldots$
- A and B are associated more strongly than would be expected by chance (α = 0.001)

Association measure evaluation

Measure of association	Correctly classified
Pointwise Mutual Information	**93.33%**
Fisher’s Exact Test	91.37%
Frequency	90.96%
None (Baseline)	80.92%

Number of features

Wrapped Progressive Sampling (van den Bosch, 2004)



- TiMBL has many parameters:
  - Nr. of nearest neighbours
  - Feature vector distance measure
  - Neighbour weighting
  - Feature weighting
- Wrapped Progressive Sampling can automatically converge to the optimal parameters for the data set

Appendix references

Van den Bosch, A. (2004). Wrapped progressive sampling search for optimizing learning algorithm parameters. In R. Verbrugge, N. Taatgen, and L. Schomaker (Eds.), *Proceedings of the 16th Belgian-Dutch Conference on Artificial Intelligence*, Groningen, The Netherlands

P. de Swart, M. Lamers, and S. Lestrade. Animacy, argument structure, and argument encoding. *Lingua*, 118(2):131–140, 2008.

N.M. Fraser and G.G. Corbett. Gender, animacy, and declensional class assignment: a unified account for Russian. *Yearbook of morphology*, 1994:123–150, 1995.