The first results of the Groot Nationaal Onderzoek Woordenschat

Marc Brysbaert, Emmanuel Keuleers, Pawel Mandera & Michael Stevens

Ghent University (Belgium)

Some developments in psycholinguistics that may be of interest to NLP researchers

1. Validation of word frequency measures
   - with lexical decision times and naming times
   - word frequencies based on film and television subtitles give the highest correlations (SUBTLEX)
2. Collection of processing times for thousands of words (megastudies)
   - English Lexicon Project
   - Dutch Lexicon Project I and II
   - French Lexicon Project

Some developments in psycholinguistics that may be of interest to NLP researchers

3. Collection of word ratings for thousands of words used in regression analyses
   - Age of acquisition (30K English words, AMT)
   - Valence and arousal (13K English words)
   - Concreteness (63K English words)
   - Same values for Dutch (30K words)
   - Investigate how well we can predict rating values on the basis of seed ratings and semantic similarity of words

A problem nagging in the background

- We need a master list of words known to the participants
  - better for our ratings
  - also necessary to calculate correct values for some measures (e.g., neighbourhood size, similarity to closest words)
  - We use word frequency as a proxy for known, but there are indications that this may not always work

Master list

- From all types of copyright-free sources (so, no existing dictionaries)
- Words ***known*** to participants and likely to be stored in their lexicon:
  - lemmas (inflections only if they are frequent)
  - pruning of derived words and compound words on the basis of:
    - frequency
    - length
    - transparency

Master list

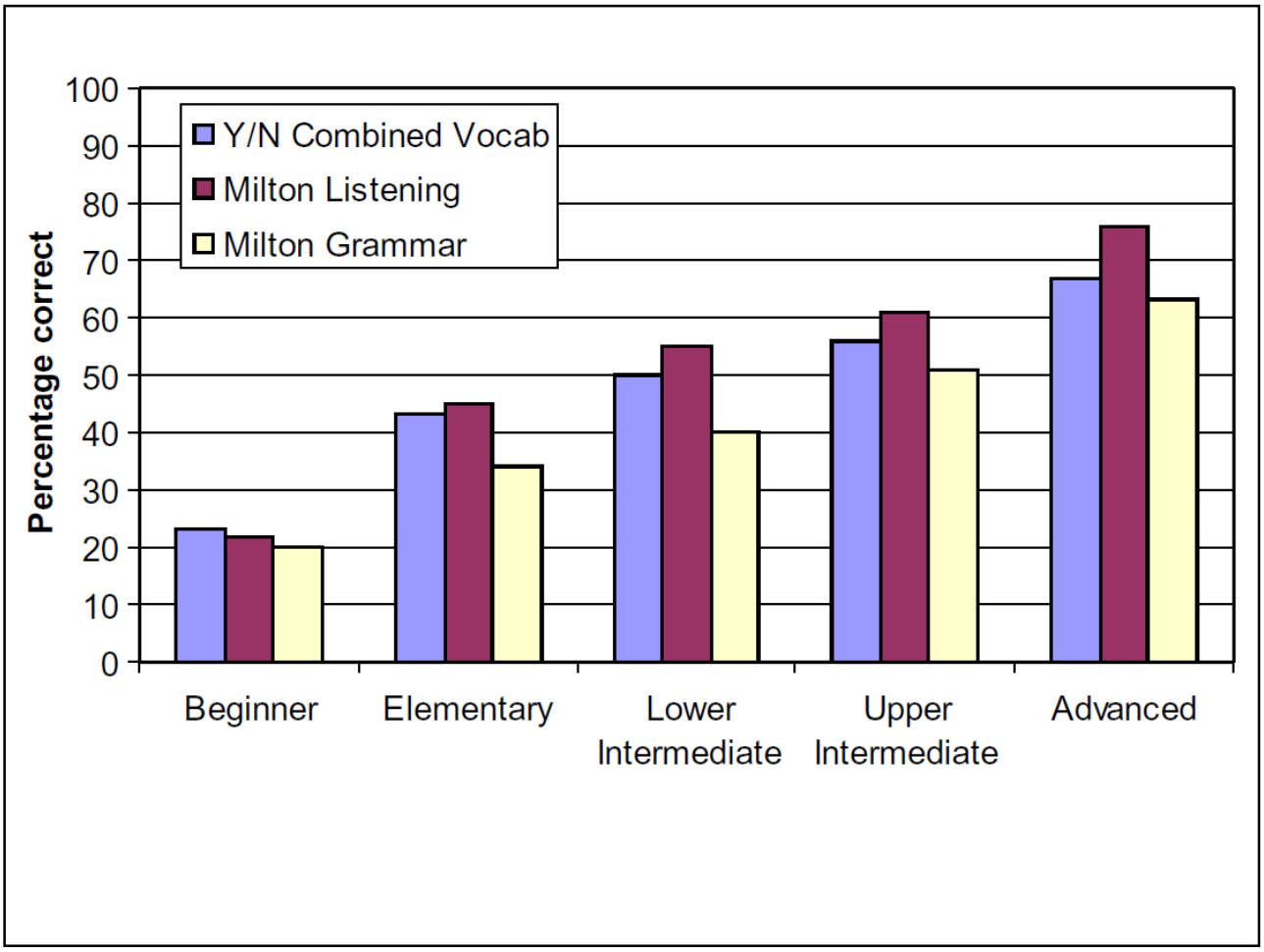
- So, ‘voetbal’, ‘voetballen’ ‘voetballer’, and ‘voetbalploeg’, but not:
  - voetbalt, voetbalspeler, voetbalbroek, voetbalgeweld, voetbalhumor, voetbalintelligentie, voetbalkoffer, ...
- Fuzzy boundaries
  - In the end element of arbitrariness which words were included and which not (preferred total size: 50-65K words)

Ways to decide which words are known

- In the rating studies participants have the option to indicate they do not know the word well enough to give a rating
- Ask participants which words they know

The Yes/No vocabulary test of L2 proficiency

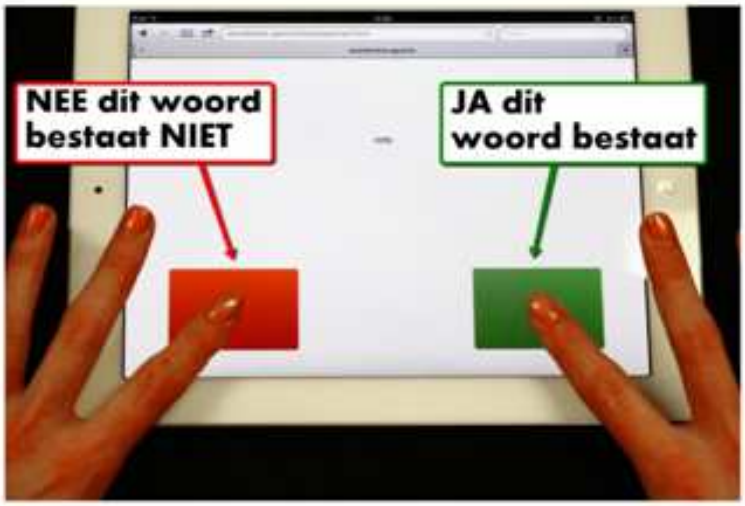
- Meara : A list with words and nonwords in which participants have to indicate the words they know is a valid test of vocabulary size
- Lemhöfer & Broersma (2011): Lextale
- Further validation studies
  - Harrington & Carey (2009)


Percentage correct
100
90
80
70
60
50
40
30
20
10
0
Y/N Combined Vocab
Milton Listening
Milton Grammar
Beginner
Elementary
Lower Intermediate
Upper Intermediate
Advanced

Groot Nationaal Onderzoek

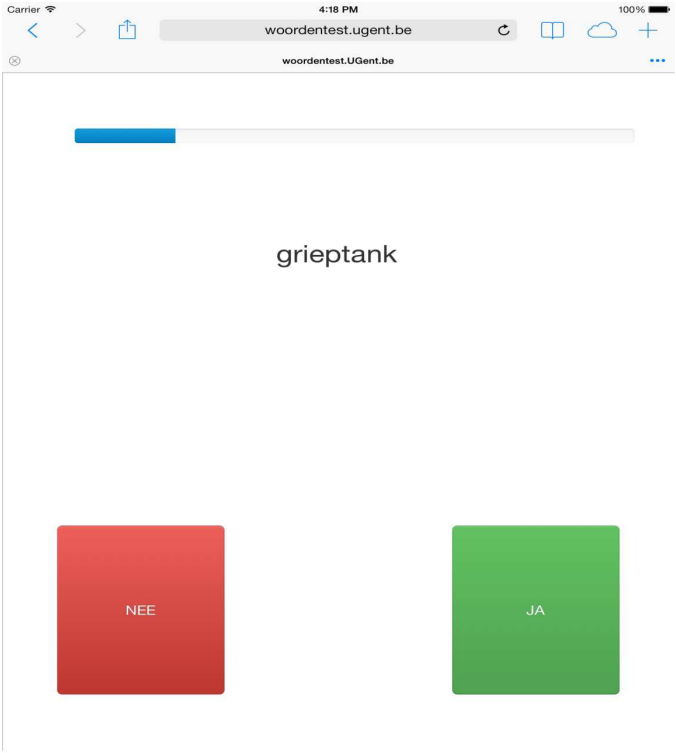
- Initiative of the Dutch broadcasters NTR and VPRO to run nationwide scientific studies (sleep, stress, perception of emotions, mathematics proficiency)
- We proposed to test vocabulary with the Y/N test
- BIG difference: we did not use a single list of 70 words and 30 nonwords, but 735 different lists
  - So, we collected data for 52,847 Dutch words and 20,653 nonwords (500K participants)


Wetenschap24
HOME NIEUWS AGENDA & GIDS KIJK & LUISTER PROGRAMMA'S SPECIALS TEST JEZELF OVER W24
Zoek naar
meer
groot nationaal onderzoek
vpro ntr:
HET GROOT NATIONAAL ONDERZOEK ONDERZOEKEN CONTACT FAQ
HOME / PROGRAMMA'S / HET GROOT NATIONAAL ONDERZOEK / ONDERZOEKEN / HET TAALONDERZOEK
Gedeeld door
272
Het Taalonderzoek
11 maart 2013
162 reacties
Laatste reacties:
Er zijn geen resultaten gevonden in de encyclopedi...
Dit is prachtig! Dank je. Zolang de cursisten aang...
Ik heb al mijn cursisten, buitenlanders die Nederl...
Meer reacties
Reageer
Wie heeft de grootste woordenschat? Nederlanders of Vlamingen? En welke woorden kennen ze wel in Groningen maar niet in Zeeland? In het vijfde Groot Nationaal Onderzoek maken we een unieke momentopname van de Nederlandse taal.
taal
Aamborstig? Of uilvorstig? Een van deze woorden bestaat niet, de ander wel. Taal is dynamisch, woorden verdwijnen en er komen nieuwe bij. Hoeveel woorden kennen Friezen wel, maar Brabanders niet? En hebben Nederlanders een grotere woordenschat dan Vlamingen? Tijd voor het Groot (Inter)Nationaal Onderzoek naar onze taal, dat moet leiden tot een staalkaart van het Nederlands in 2013. Nooit eerder is in de wereld een dergelijk onderzoek naar taal gedaan.
Het Taalonderzoek is het vijfde Groot Nationaal Onderzoek, een initiatief van NTR, VPRO en de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO). Onder leiding van taalonderzoekers van de Universiteit van Gent is een test ontwikkeld om te meten hoeveel woorden iemand
Lees ook:
NEDERLAND - BELGIË
74 72
De resultaten van het GNO Taal
Nederlanders hebben de grootste woordenschat.
Kijk ook: Labyrint TV
Groot Nationaal Onderzoek Taal
Marc Brysbaert en



In deze test krijg je 100 woorden te zien, zowel bestaande als niet-bestaande woorden.
Geef voor elk woord aan of het volgens jou een ***Nederlands woord is, of niet.***
Met een Nederlands woord bedoelen we dat het woord in (een deel van) het Nederlandse taalgebied gebruikt en begrepen wordt.

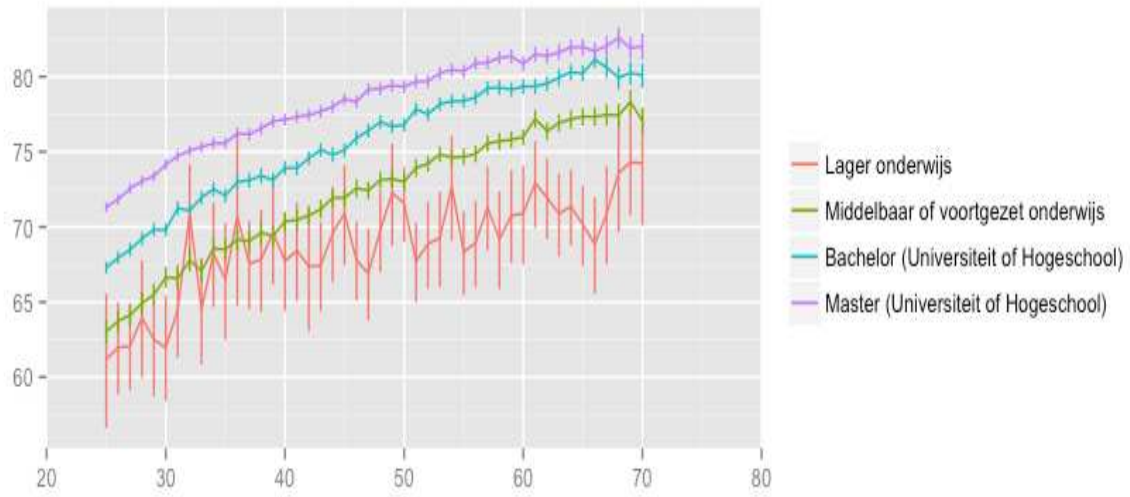
http://woordentest.ugent.be


Carrier
4:18 PM
100%
woordentest.ugent.be
woordentest.UGent.be
grieptank
NEE
JA


Jouw resultaat
Op basis van je resultaten schatten we dat je 86% van alle Nederlandse woorden kent.
Deel je score: Facebook 851 Twitter 25 Google+ Email
Je hebt 99% van de woorden juist herkend.
Je hebt 13% van de niet-bestaande woorden verkeerd herkend.
Dit geeft je een gecorrigeerde score van 99% - 13% = 86%.
Hiermee behoor je tot de absolute topgroep!
Verbeter je score
Wat betekent dit?
Doe ook de auteurstest
Bekijk je antwoorden opnieuw
Bestaande woorden die je niet herkende
Niet-bestaande woorden die je als bestaand woord herkende
Bestaande woorden die je wel herkende

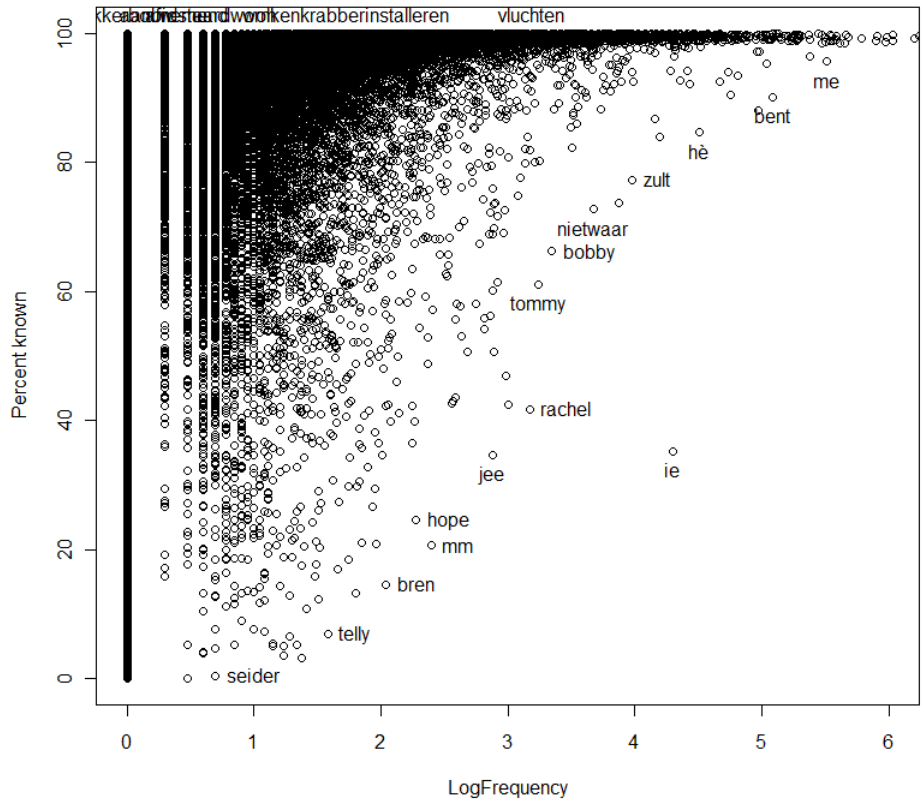
Two types of analyses possible

1. Per participant
   - Quite easy to correct for response bias: % yes to words minus % yes to nonwords
   - Allows us to investigate individual differences


80
75
70
65
60
20
30
40
50
60
70
80
Lager onderwijs
Middelbaar of voortgezet onderwijs
Bachelor (Universiteit of Hogeschool)
Master (Universiteit of Hogeschool)

Two types of analyses possible

2. Per word
   - The information we were after
   - Challenge here is how to best correct for guessing
   - Still, even with some simple correction you get very interesting results
     - e.g., words known in the North/South but not known to the other community
     - relationship between frequency (SUBTLEX) and percentage known


Percent known
LogFrequency
0
20
40
60
80
100
0
1
2
3
4
5
6
vluchten
me
bent
hè
zult
nietwaar
bobby
tommy
rachel
jee
ie
hope
mm
bren
telly
seider

r = .50

Further exceptions

- 22K words not in SUBTLEX, of which 11K known to more than 75% of the participants
  - unavailable for psycholinguistic research based on the word SUBTLEX frequency list alone!
- Words not in SUBTLEX corpus and known to everyone:
  - akkerbouw, baanbreker, bestuiving, bouwgrond, deelwoord, flitspaal, globaliseren, gospelmuziek, hamsteraar, kijkcijferkanon, oppositiepartijen, overheidstaken, postpakket, proeflokaal, puntbaard, ramptoerist, rechtsbeginsel, regeerakkoord, scheurkalender

Further exceptions

- Words twice in SUBTLEX corpus and known to everyone:
  - aanbidster, verfpot, zwerftocht, klapstoel, regenwolk, kaasplank, schietgebed, dorpsgenoot, trekvogel, graanproduct, bierglas, inleidend, smaakstof, kernwoord, kortharig

Other uses of the data

- Programs to estimate difficulty level of texts
- Dictionaries (e.g., some words no longer used, words to be included in translation dictionaries)
- Distribution of word knowledge
- Research into the recognition of compound words and derived words
- ...
- freely available at http://crr.ugent.be/