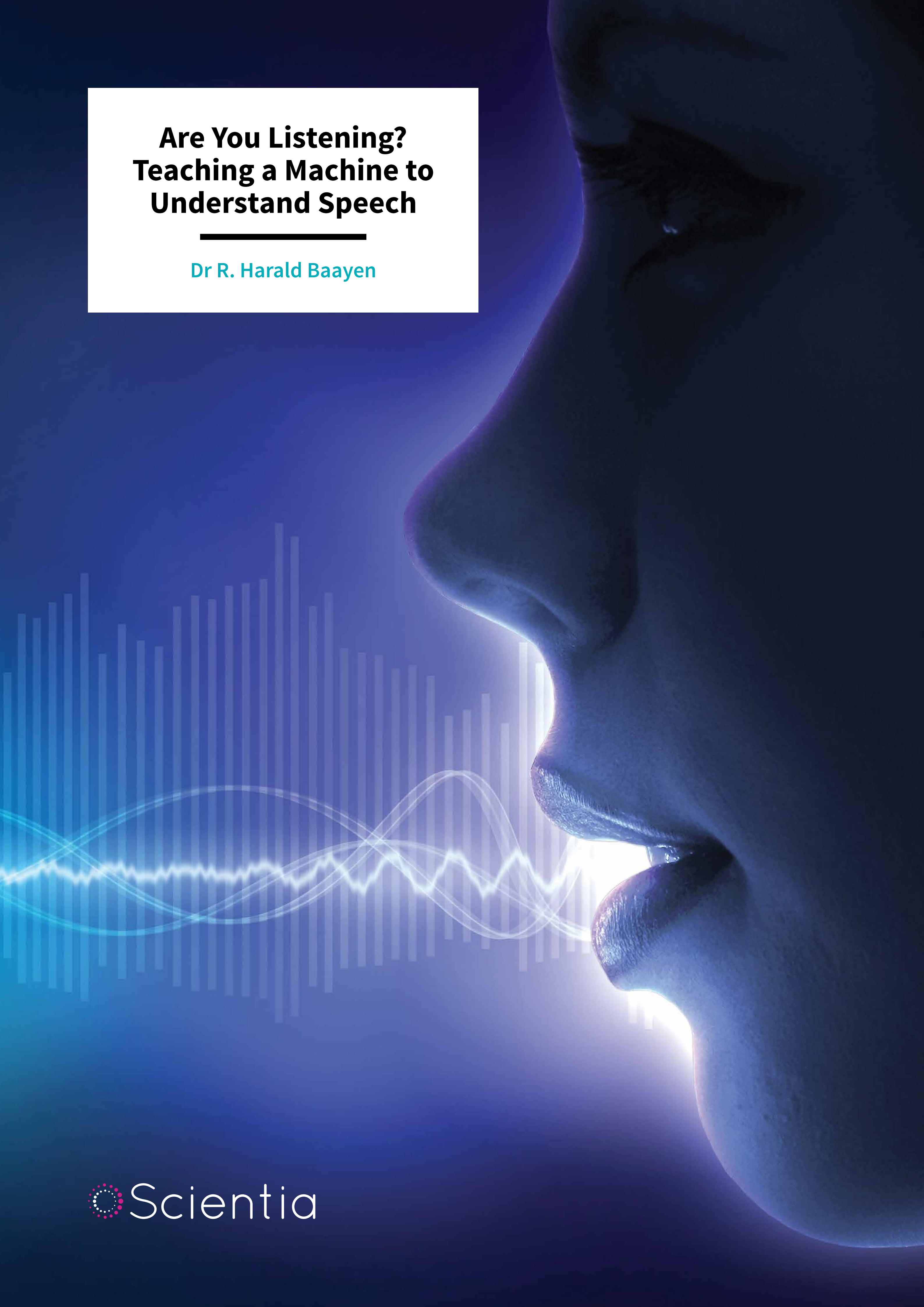
Are You Listening? Teaching a Machine to Understand Speech

Dr R. Harald Baayen



Scientia

ARE YOU LISTENING? TEACHING A MACHINE TO UNDERSTAND SPEECH

In the past few years, speech recognition has become a new standard for state-of-the-art technology. We now talk to our phones as much as we talk on them. How can helping machines learn to listen improve our understanding of how our own brains work? **Dr Harald Baayen** at Eberhard Karls University Tübingen and his collaborators work at the intersection of linguistics, psychology, and computational data science to illuminate elegant solutions for processing speech.

What's So Hard About Speech?

We've all been there. You say something to Siri or Alexa or Google, and what she repeats back is a baffling far cry from your original statement.

'Hey Siri, what's the weather in Denver today?'

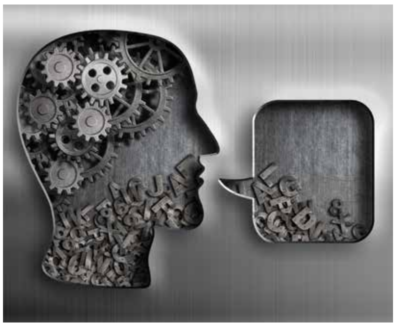
'Ok. Calling Mom.'

Why do even our most advanced versions of speech recognition software struggle so much with understanding simple requests? It turns out that listening isn't as simple as it might seem. When your digital assistant is trying to figure out what you just said, it has a lot to parse out. The unique way you pronounce the letter O, the subtle shift between 'then' and 'than', the two different ways you say the 're' in 'record' when talking about recording a record. Scientists are just beginning to scratch the surface of how the brain processes speech and how we might train computers to do the same.

Breaking Down Words

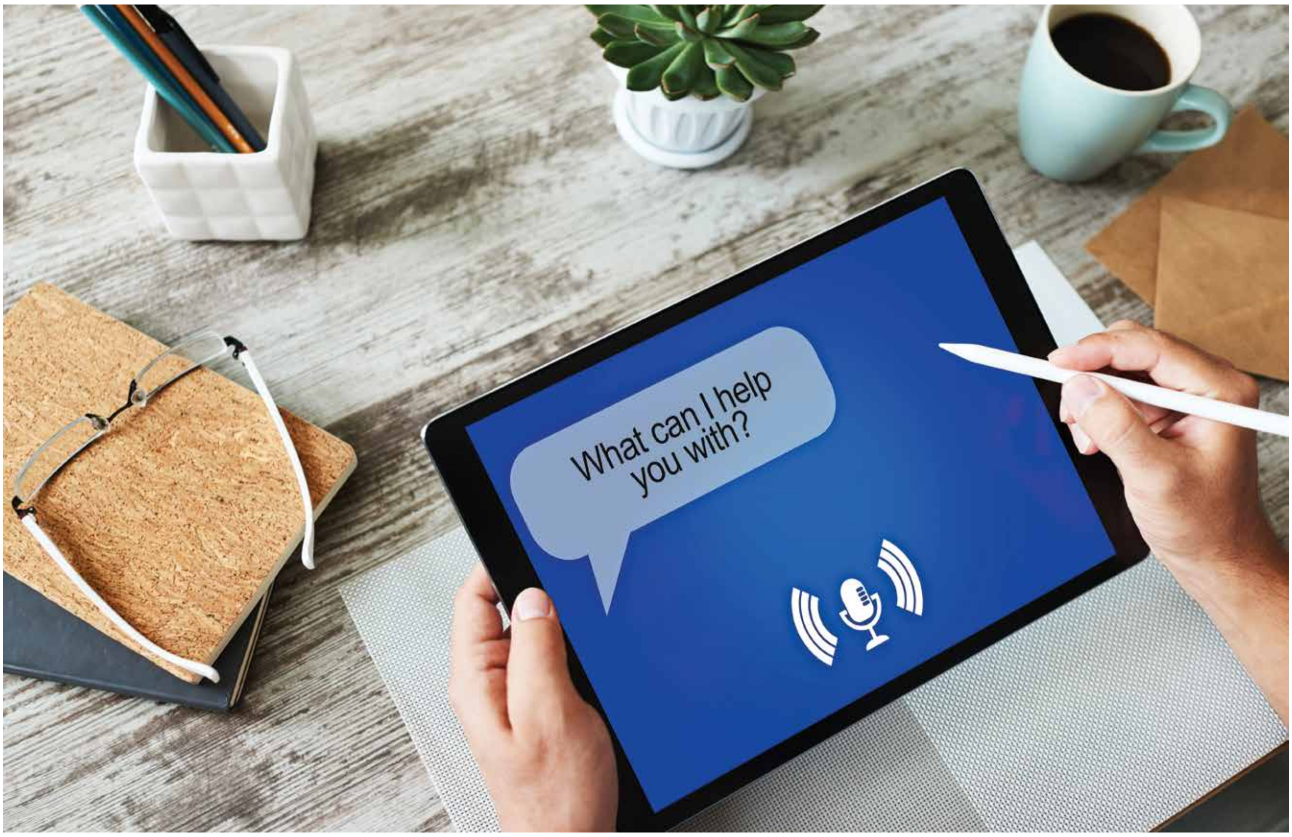
In introductory linguistics, students learn that words can be broken down into various units depending on what you are aiming to study. Phonemes are the units of sound that compose words in a particular language. For example, the word 'at' has two sounds, 'ah' and 't', which are both phonemes in English. Lexemes are units of meaning, a word and all of its related forms. For example, the lexeme 'jump' includes the forms 'jumped', 'jumps', 'jumping', all of which indicate that someone or something leapt or will leap into the air.

Morphemes are the smallest units of words that still carry meaning. Many of the morphemes in a language are words themselves, but prefixes and suffixes that modify the meaning of a word can also be morphemes. For example, the word 'dogs' contains two morphemes: the word 'dog', which conveys a domestic canine, and the word modifier 's', which means there's more than one of them.



Many studies have posited that morphemes are natural units of all languages, with evidence for morpheme processing areas in the brain. Some schools in linguistics focus on the morpheme as an important component of language processing. Many computational models of language processing focus on morphemes as critical units, creating complex systems to parse meaning out of complex words that the human brain identifies with ease. These models typically require a lot of manual input – researchers need to set up complex rule systems, define morphemes and their alternative forms, devise mechanisms for dealing with exceptions, and give up on irregular forms that are not well predictable by rules.

Computational psycholinguist, Dr Harald Baayen of Eberhard Karls University Tübingen and his colleagues recognised that there had to be a better



way. Utilising a much simpler algorithm, dubbed the Linear Discriminative Learner (LDL), the research team focuses on modelling triphones (sequences of three phonemes) rather than morphemes. Within a given language, triphones are the possible combinations of three sounds. For example, in English, 'kin' ('k-ih-n') is a possible triphone, but 'kni' ('k-n-ih') is not. It is important to note here that triphones focus on sounds and not spelling – the English word 'knife' consists of the triphone 'n-eye-f'. An LDL of English words would include 'kin' but not 'kni'. LDL is not concerned with the morphemes present in a word, just its sound components.

'A crucial part of my model is that form and meaning are represented by numerical semantic vectors, and simple linear transformations between these vectors turn out to work surprisingly well for modelling comprehension and production,' explains Dr Baayen. In his model, semantic vectors quantify to what extent a given word makes one think of any other word. 'Ship', for example, might make one think primarily of 'sea', 'captain', 'cargo' and 'waves', whereas 'pasta' may lead to thoughts of 'Italy', 'spaghetti', 'pesto' and 'tomato sauce'.

The team's LDL model yields surprisingly elegant results. With relatively simple calculations, it is able to map word form to meaning, and vice versa, without hundreds of rules for morpheme forms necessary for a typical language processing program. Further, it may provide evidence that morphemes aren't as natural a component of language as thought. Triphone maps naturally organise into clusters similar to morphemes, without the complex computations needed to generate the same maps from morphemes. Language processing that focuses on sound patterns rather than whole words offers a simpler way to get to the same solution. Dr Baayen comments, 'One thing that is becoming increasingly clear is that these classical symbolic approaches severely underestimate how very rich our speech is, and how informative this richness is.'

Tracking Speech Sounds

Sounds could be a simpler focus for the machine processing of speech. However, relying on sounds can be tricky too. Frequently used words tend to be shortened in speech, to form what linguists call 'reductions'. Syllables get dropped, vowels get shortened, and the words in commonly used phrases begin to blend together. If you've ever said 'ain't' or told anyone to 'c'mon', you've participated in this phenomenon. 'One of the most striking aspects of reductions is that if you listen to them out of context, you have no idea what is meant,' notes Dr Baayen. Understanding reductions is essential for creating computer programs that can recognise the words in a sentence correctly.

Dr Baayen and his colleagues recognised that multiple factors are at play when predicting how a person will say a given word. Theories of why common words get shortened often focus on reduced effort but miss a critical point: most people have more practice saying common words. How many times have you said the word 'the' in your life? The team predicted that words we say frequently may be shorter because we are better practiced in the mouth movements necessary to form them in different contexts. For example,

the muscles in your mouth and tongue must go through different transitions to say ‘the’ when you’re saying ‘touch the screen’ versus ‘to the moon’. However, since you have so much practice saying ‘the’, it is likely that your mouth can rely on muscle memory to quickly say either phrase.

Dr Baayen and his collaborators tested this prediction in an experiment with native German speakers. Participants were recorded saying similar words that vary in frequency in the German language. The team focused on two factors, clarity and smoothness. Clarity indicates that syllables were spoken correctly, while smoothness indicates that sounds were shifted to make the transitions between words easier. The researchers found that the vowels of infrequent words were spoken clearly, with a lowered jaw, and that for medium frequency words the jaw was lowered less, with somewhat reduced tongue movements favouring smoothness. Interestingly, high-frequency words were able to cater to both clarity and smoothness, not lowering the jaw by much but at the same time articulating with large articulatory gestures. The researchers posit that the practice in saying these words allows speakers to more quickly transition the muscles of the mouth without losing clarity.

It is not only the frequency of a word that influences how it is said. Even the same word said by the same person can sound quite different across situations. For example, think of the difference in your speech when you are very excited versus when you are very tired. The rate, annunciation, and pronunciation of each phoneme in your words would vary dramatically, even if you were speaking the same sentence. Despite this variance, we typically have no problem understanding words that have shortened or alternatively pronounced phonemes – we often don’t even notice.

A Simple Solution

Historically, most linguistic theories have assumed that speech comprehension is phoneme based, requiring complex neural networks to interpret variations in phoneme pronunciation. Dr Baayen and his colleagues have developed computational speech recognition algorithms that recognise speech within the human range of accuracy, without using phonemes at all. Instead, the program focuses on changes in sound frequency with each word, using a simple but wide network that has numerous acoustic features as inputs. The success of the program offers hope for advanced iterations of digital assistants that can understand speech better without the need for complex processing systems. ‘We are exploring what can be done with much simpler networks, but networks with lots of units, hence “WIDE” learning networks, combined with smart features.’

In an expansion of this work, Dr Baayen and his colleagues have developed a computational network that can recognise isolated words with greater accuracy than many of the more complex speech recognition programs on the market today. Using acoustic features that summarise patterns of change in the different frequency bands that the cochlea in the human ear is sensitive to, the native discriminatory learning (NDL) system was trained to learn words by watching hours of TV news broadcast. Using a layered wide network framework, the system excels at recognising single words without the computational complexity of standard speech processing software. Further, it is capable of improving accuracy the longer it learns from a speaker. The elegance of the team’s findings suggests that neural networks for speech recognition in the brain could be far simpler than previously thought. However, the team has more work to do, as they have not yet shown that their system can be expanded to understanding words in running speech.

Appreciating the Experienced Mind

The team’s speech recognition work ties research from the fields of linguistics, psychology, and computational data science to both improve technology and illuminate how neural networks in the brain may work. His experience in these arenas has led him to investigations of another aspect of the human experience – aging.

It is commonly held that aging is associated with cognitive decline. Older individuals often demonstrate slower reaction times during memory tests, that is assumed to be the result of neuron deterioration in the brain. Dr Baayen and his collaborators have come to a different conclusion – these slower times are not the result of decline, but of the limits of information processing speeds in the brain. Older adults have accumulated a lifetime of knowledge that they must sort through to get to a particular fact. It takes longer for the brain to search these vast memory stores, slowing reaction times. Dr Baayen and his colleagues have harnessed their knowledge of the brain and computational prowess to build simulations of this phenomenon that predict performance results seen in real life.

The team’s research is shifting our view of how the brain handles information, accelerating our understanding of how speech is processed, and improving our technology’s ability to understand us.

Meet the researcher

Dr R. Harald Baayen
Department of Linguistics
Eberhard Karls University Tübingen
Tübingen
Germany

Dr Harald Baayen began his linguistics career at the Free University of Amsterdam, obtaining a bachelors, masters, and PhD in General Linguistics. He went on to join the scientific staff at the prestigious Max-Plank Institute for Psycholinguistics for eight years, before joining the faculty at the University of Alberta, Edmonton and then Eberhard Karls University, Tübingen as a Professor of Quantitative Linguistics. His work focuses on human speech, both how it is generated and how it is processed by the brain and machines.

CONTACT

E: harald.baayen@uni-tuebingen.de
W: http://www.sfs.uni-tuebingen.de/~hbaayen/contact.html

KEY COLLABORATORS

Michael Ramscar (aging), Eberhard Karls University, Tübingen
Elnaz Shafaei (auditory comprehension), Eberhard Karls University, Tübingen
Fabian Tomaschek (articulography), Eberhard Karls University, Tübingen
Yu-Ying Chuang (NDL modelling), Eberhard Karls University, Tübingen

FUNDING

Alexander von Humboldt Professorship (2011–2016)
DFG research unit (2016–)
ERC advanced grant (2017–)

FURTHER READING

RH Baayen, YY Chuang, JP and Blevins, Inflectional morphology with linear mappings, The Mental Lexicon, 2018, 13, 232–270.

F Tomaschek, BV Tucker, M Fasiolo, and RH Baayen, Practice makes perfect: The consequences of lexical proficiency for articulation, Linguistics Vanguard, 2018, 4, 1–13.

D Arnold, F Tomaschek, K Sering, F Lopez, and RH Baayen, Words from spontaneous conversational speech can be recognized with human-like accuracy by an error-driven learning algorithm that discriminates between meanings straight from smart acoustic features, bypassing the phoneme as recognition unit, PLoS ONE, 2017, 12, e0174623.

M Ramscar, CC Sun, P Hendrix, and RH Baayen, The Mismeasurement Of Mind: Lifespan Changes in Paired Associate Test Scores Reflect The 'Cost' Of Learning, Not Cognitive Decline, Psychological Science, 2017, 28, 1171–1179.

RH Baayen, C Shaoul, J Willits, and M Ramscar, Comprehension without segmentation: A proof of concept with naive discrimination learning, Language, Cognition, and Neuroscience, 2015, 31, 106–128.

M Ramscar and RH Baayen, The myth of cognitive decline: why our minds improve as we age, New Scientist, 2014, 221, 28–29.

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN