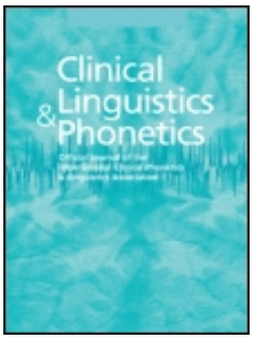

Clinical Linguistics & Phonetics

ISSN: 0269-9206 (Print) 1464-5076 (Online) Journal homepage: https://www.tandfonline.com/loi/iclp20

Norwegian Words: A lexical database for clinicians and researchers

Marianne Lind, Hanne Gram Simonsen, Pernille Hansen, Elisabeth Holm & Bjørn-Helge Mevik

To cite this article: Marianne Lind, Hanne Gram Simonsen, Pernille Hansen, Elisabeth Holm & Bjørn-Helge Mevik (2015) Norwegian Words: A lexical database for clinicians and researchers, Clinical Linguistics & Phonetics, 29:4, 276-290, DOI: 10.3109/02699206.2014.999952

To link to this article: https://doi.org/10.3109/02699206.2014.999952

Published online: 14 Jan 2015.

Clinical Linguistics & Phonetics, April 2015; 29(4): 276–290

ISSN: 0269-9206 print / 1464-5076 online
DOI: 10.3109/02699206.2014.999952

informa
healthcare

Norwegian Words: A lexical database for clinicians and researchers

MARIANNE LIND[1,2], HANNE GRAM SIMONSEN[1], PERNILLE HANSEN[1], ELISABETH HOLM[1], & BJØRN-HELGE MEVIK[3]

[1]*Department of Linguistics and Scandinavian Studies, MultiLing (CoE), University of Oslo, Oslo, Norway,* [2]*Department of Speech and Language Disorders, Statped, Oslo, Norway, and* [3]*Department for Research Computing, University Centre for Information Technology, University of Oslo, Oslo, Norway*

(*Received 15 September 2014; revised 11 December 2014; accepted 15 December 2014*)

Abstract
All words have properties linked to form, meaning and usage patterns which influence how easily they are accessed from the mental lexicon in language production, perception and comprehension. Examples of such properties are imageability, phonological and morphological complexity, word class, argument structure, frequency of use and age of acquisition. Due to linguistic and cultural variation the properties and the values associated with them differ across languages. Hence, for research as well as clinical purposes, language specific information on lexical properties is needed. To meet this need, an electronically searchable lexical database with more than 1600 Norwegian words coded for more than 12 different properties has been established. This article presents the content and structure of the database as well as the search options available in the interface. Finally, it briefly describes some of the ways in which the database can be used in research, clinical practice and teaching.

Keywords: Age of acquisition, frequency, imageability, language processing, phonological neighbourhood density

Introduction

Words and other lexical items have inherent properties which influence how easily they are accessed from the mental lexicon in language production, perception and comprehension. Examples of such properties are conceptual features like imageability and concreteness, linguistic features like phonological and morphological complexity, word class, inflection class and argument structure, and usage-related features like frequency of use and age of acquisition.

These word properties are universal in the sense that all languages have words, and all words have properties linked to form, meaning and usage patterns. However, since languages and cultures differ, the actual properties associated with a particular word in a given language are unique to that word. Even though the word can be translated to another language, and several of

Correspondence: Marianne Lind, Department of Linguistics and Scandinavian Studies, MultiLing (CoE), University of Oslo, P.O. Box 1102 Blindern, N-0317 Oslo, Norway. Tel: +47 90936031. E-mail: marianne.lind@statped.no

the properties may be similar, the two words do not necessarily exhibit exactly the same properties (Pavlenko, 2009).

An example is the word *tran* in Norwegian, which translates as *cod liver oil* in English. Both of these words are nouns, but structurally, the Norwegian word is phonologically and morphologically simpler than its English counterpart. The concept is lexicalised as a short single word in Norwegian, whereas it is expressed as a longer compound in English. Furthermore, *tran* is rated as highly imageable by Norwegians, probably due to the fact that many Norwegian children (and adults as well) swallow a spoonful of cod liver oil every day, at least through the winter months. Our guess would be that in general *cod liver oil* would get a lower imageability rating by native speakers of English (Simonsen, Lind, Hansen, Holm, & Mevik, 2013). Since word properties differ between languages, we need language specific information about the various properties if we are going to take them into account in research and clinical practice in different linguistic communities. Such language specific information is also necessary for cross-linguistic and bilingual research.

Several of the properties of a word, such as word class, phonological and morphological complexity, and argument structure, can be established by consulting a dictionary or a reference grammar or by relying on the native linguistic knowledge and intuition of the researcher/clinician. For other properties, such as imageability and frequency, there is no established consensus on the values assigned to a given word. Rather, the values depend on language use and personal experience. For such properties, the values for single words must be established as mean values based on ratings or usage patterns collected from many individuals.

In order to facilitate research on word processing as well as development of assessment tools and treatment tasks for speech and language therapy, a lexical database with more than 1600 words coded for a number of different properties has been established for Norwegian Bokmål.[1] The database, which was launched in 2013, is called *Ordforrådet*.[2] It is available for free from the website of The Text Laboratory at the University of Oslo (http://tekstlab.uio.no/ordforradet/en). In 2014, an English interface was added, allowing for easier access to the database for researchers who would like to conduct cross-linguistic research on Norwegian and one or more other language(s). In the English interface, the database is called *Norwegian Words*, and in the following, we will refer to the database by its English name. In this interface, English translations of the words in the database are provided, and all instructions and explanations are given in English. The actual words in the database and the property values are of course identical in the Norwegian and the English interfaces.

As an online resource and tool *Norwegian Words* is not unique. There are other collections of words coded for one or more (psycho)linguistic properties, most notably, the *MRC Psycholinguistic Database* (http://www.psych.rl.ac.uk; Wilson, 1988), which contains more than 150 000 words and 26 linguistic and psycholinguistic properties. Despite the fact that not all the words are coded for all the properties, this database is an impressive resource. However, it is entirely and solely a database for English, and no matter how useful it may be for researchers and clinicians working on English language data, it cannot automatically be transferred to similar work on other languages.

In this article, we present the content and structure of the database *Norwegian Words* as well as the search options available in the interface. To get a fuller picture of what the database looks like and how it works, we recommend a visit to the website. By describing the methodology that we used when establishing this database, we hope to inspire others who would like to establish similar

[1]In Norwegian, there are two official written standards (Bokmål and Nynorsk) as well as a number of dialects that are widely used in formal as well as in less formal settings. The written standards and most of the dialects are mutually intelligible. Bokmål is used as the written language by the majority of Norwegians.

[2]Literally, *ordforrådet* translates as *the vocabulary.*

kinds of tools for research and clinical practice in their language(s). Finally, we briefly present and discuss some of the ways in which *Norwegian Words* has been used and can be used in research, teaching and clinical practice and look into some of the possibilities for further development and use of the database.

Words and properties in the database

Selection of words

Norwegian Words contains 1651 words: 917 nouns, 509 verbs and 225 adjectives. All the words belong to open word classes, and no function words are included. In order to ensure clinical relevance of the database, the words were mainly selected from various assessment tools for language acquisition and language disorders in Norwegian. Words from the following tests and assessment tools were included:

- The Norwegian version of *MacArthur-Bates Communication Development Inventories* (MB-CDI; Kristoffersen & Simonsen, 2012; Simonsen, Kristoffersen, Bleses, Wehberg, & Jørgensen, 2014). Approximately 460 nouns, verbs and adjectives from the parental checklists that were used to collect CDI-data for Norwegian are included in the database. Only words that exist in adult Norwegian are included. For instance, sound effects, games and routines have been excluded.
- The Norwegian version of the *Verb and Sentence Test* (VAST), which includes the *Past Tense Test* (Bastiaanse, Lind, Moen, & Simonsen, 2006; Ragnarsdóttir, Simonsen, & Plunkett, 1999). Approximately 160 verbs (target verbs and distractors) from this test battery are included in the database.
- The Norwegian version of the *Psycholinguistic Assessments of Language Processing in Aphasia* (PALPA; Kay, Lesser, & Coltheart, 2009). Approximately 400 words, mostly nouns, but also some verbs, from PALPA are included in the database.
- A list of 300 ''best words'' constituting the basis for cross-linguistic lexical tasks (CLTs) for language assessment of multilingual children (Haman, Łuniewska, & Pomiechowska, in press). The Norwegian version of this list (i.e. the target words for the Norwegian CLT; Simonsen, Hansen, & Łuniewska, 2012) is included in the database.

In addition, verbs, nouns and adjectives were extracted from a corpus of transcribed semi-spontaneous narratives (oral picture descriptions) by 40 speakers with aphasia and 60 neurologically healthy speakers (Korpijaakko-Huuhka & Lind, 2012; Lind, Kristoffersen, Moen, & Simonsen, 2009). A few verbs and nouns were also added that were not included in the assessment tools or the narratives. These were verbs and nouns for which a homonym noun or verb already existed in our data. For instance, since the verb *å danse* ''to dance'' was represented in our data, we added the noun *en dans* ''a dance''. This addition of noun and verb homonyms facilitates exploration of possible effects of what is referred to as a ''name relation between a noun and a verb'' (Jonkers & Bastiaanse, 2007) on word production and comprehension in individuals and different groups of speakers (cf. also Kambanaros, 2013). There are in total 350 nouns and verbs in *Norwegian Words* for which a homonym from the opposite word class also exist in the database. Since a single word may have more than one homonym (e.g. the verb *å sparke* ''to kick'' and the two nouns *et spark* ''a kick'' and *en spark* ''a kicksled''; a Nordic vehicle propelled like a kick scooter), these 350 words make up more than 175 pairs.

In the database, the words are treated as lemmas, meaning that only one form exists for each word. For instance, there are not separate entries for different inflectional forms of a word. Verbs are presented in the infinitive, with the infinitive marker (*å hoppe* ''to jump''), and nouns are

Table 1. Properties and property values of the words in *Norwegian Words*.

Properties	Values
Word class	Noun; verb; adjective
Word length	Number of letters; number of phonemes; number of syllables
Morphological complexity	Derivation; compound
Phonological complexity	Word initial fricative; word initial consonant cluster
Frequency of use	Low; medium; high
Phonological neighbourhood density	Few; some; many
Imageability	Low; medium; high
Subjective age of acquisition	Early; medium; late

mostly presented in the indefinite singular form with the indefinite article (*en gutt* ''a boy''). Nouns that can be interpreted both as mass nouns and count nouns, are given two entries, one with the indefinite article (*et vann* ''a lake'') and one without (*vann* ''water''). A certain level of polysemy among the words in such a database is unavoidable, but homographs which cannot be disambiguated with an infinitive marker or an indefinite article (e.g. *å sprette,* which can mean both ''to bounce'' and ''to open'' (e.g. a bottle)), were excluded.[3]

In Norwegian, there is no clear difference between adjectives and manner adverbs. Some of the words that are classified as adjectives in our database are frequently used as adverbs (*fort* ''quick/quickly''), whereas others are predominately used as adjectives (such as colour terms). For simplicity, they are all categorised as adjectives in the database. Among the adjectives there are 12 present participles (e.g. *utdypende* ''detailed'' from *utdype* ''(to) detail (elaborate)'') and 19 past participles (e.g. *irritert* ''annoyed'' from *irritere* ''(to) annoy''). All the words are coded for the set of properties presented in Table 1.

In addition to the properties listed in Table 1, the nouns and verbs in the database are coded for certain word class specific properties. The nouns are classified according to grammatical gender. In Norwegian, there are three gender classes: masculine, feminine and neuter. As a rule, each noun belongs to one class, and the gender of the head noun determines the form of the determiner and modifiers in the noun phrase. Furthermore, the nouns are classified as countable or mass nouns, and nouns that are very rarely used in the singular (e.g. *briller* ''glasses'', *penger* ''money''), are coded with the information ''mainly used in the plural''. For the verbs in the database, information is available on transitivity and valency (the number of obligatory arguments). In addition, one can choose to access single word verbs (e.g. *å gå* ''to go/walk'') or phrasal verbs consisting of more than one word (e.g. *å få til* ''to manage'').

As mentioned, the values of several of the properties listed in Table 1 can be established by consulting dictionaries, grammars or the linguistic intuitions of a few native speakers. This is true for word class, morphological and phonological complexity, word length and the word class specific properties mentioned above. In principle, it is also true for phonological neighbourhood density, although for this property it is hardly realistic to establish the values without taking advantage of computer based dictionaries. To establish values for the rest of the properties included in *Norwegian Words* (frequency of use, subjective age of acquisition and imageability), data must be gathered from a substantial cohort of language users. In the next section, we present each of the properties listed in Table 1, with particular emphasis on phonological neighbourhood density, subjective age of acquisition and imageability.

[3]Thirteen homographs from the CLT ''best words'' list were kept, though, since they were important for cross-linguistic comparison with more than 30 other languages. These words were disambiguated with an extra word in parentheses. For instance, *et bein* translates to ''bone'' and to ''leg'', and was disambiguated by *knokkel* ''bone'' and *kroppsdel* ''body part'', respectively.

Word class, word length, morphological and phonological complexity

Each word in the database is coded for word class (noun, verb or adjective), and for each word three measures of word length are included: the number of letters, the number of phonemes and the number of syllables. Regarding morphological complexity, words that are compounds (e.g. *en kjøkkenhage* ''a kitchen garden'', *en leppestift* ''a lipstick'') and words that are derivations (e.g. *bråkete* ''noisy'', *et forræderi* ''a betrayal'') are coded as such. Phonologically, two types of structures are coded, namely words that start with a consonant cluster (e.g. *å blunke* ''to wink'', *bred* ''wide'') and words that start with a fricative (e.g. *å falle* ''to fall'', *saus* ''sauce''). These factors are known to influence language acquisition in normally developing children (Simonsen, 1990) as well as in children with phonological disorders (Grunwell, 1981).

The coding of word class and morphological and phonological complexity was conducted by several native speakers of Norwegian with a professional background in linguistics. First an MA-student of linguistics coded all the words for these properties. The categorisations were checked by at least one other native speaker/linguist. This task was performed by the first four authors of the present article who divided the words between them. In cases of doubt or disagreement concerning word class categorisation and the coding of morphological complexity, *Bokmålsordboka*[4], the main reference dictionary of Norwegian Bokmål, was used as the reference point. Word length in letters was counted automatically, and word length in phonemes and syllables was calculated automatically based on *NorKompLeks* (Nordgård, 1998), a computational lexicon for Norwegian Bokmål and Nynorsk, developed at the Norwegian University of Science and Technology. The Bokmål section, which our calculations are based on, is a phonologically transcribed version of *Bokmålsordboka*.

Frequency of use

Frequency of use is a factor which is known to influence language acquisition and word retrieval in adults. For instance, Goodman, Dale, and Li (2008) divided the words in children's vocabularies into six broad lexical categories (common nouns, verbs, adjectives, people, closed class words and other), and reported that within each of these categories, frequent words were acquired first. The effect of frequency on word retrieval in adults has been demonstrated repeatedly over many years for neurologically healthy speakers (Jescheniak & Levelt, 1994) as well as for speakers with aphasia (Kittredge, Dell, Verkuilen, & Schwartz, 2008).

In *Norwegian Words* information about the lemma frequency of each word is based on NoWaC (Guevara, 2010), a 700 million word corpus of texts from the ''no-domain'' on the Internet. Each word in the database is categorised as low, medium or high in frequency. For this categorisation, we divided all the words in the database in three groups: The 25% of the words with the lowest frequency values were classified as low frequency, the 25% of the words with the highest frequency values were classified as high frequency, and the rest of the words (the 50% in the middle) were classified as medium in frequency. For instance, the adjective *forarget* ''indignant, annoyed'' occurs 305 times in the 700 million word corpus and is a low frequency word. An example of a high frequency word is the noun *ei bok* ''a book'' which occurs more than 250 000 times in NoWaC. This classification of words as low, medium or high in frequency is of course more or less arbitrary, and users of the database may want to draw the lines differently, e.g. by dividing into high and low within each word class. To facilitate for such recategorisation, there is also information about the actual frequency value (based on NoWaC) for each word in the database.

[4]http://www.nob-ordbok.uio.no.

Phonological neighbourhood density

Phonological neighbourhood density (PND) is a measure of how common the segmental phonological structure of a word is in a language. A word comes from a phonologically dense neighbourhood if there are many other words in the language that sounds almost like it. Words that differ in one phoneme only, are phonological neighbours (Vitevitch & Luce, 1999). The difference may occur due to substitution (as in the minimal pair *wing – ring*) or addition/deletion (as in the word pair *swing – wing*). In Norwegian, vowel length and lexical tone are phonologically distinguishing features (Kristoffersen, 2000). Thus, words which differ only in vowel length (e.g. *mate* /ma:te/ ''feed'' versus *matte*/mate/''rug'') are counted as phonological neighbours. Likewise, words differing only in lexical tone (e.g. *bønder* (/[1]bøner/ ''farmers'' versus bønner/[2]bøner/''beans'') are phonological neighbours (Ribu, 2012). Words that have many phonological neighbours, for instance *et bord* (''a table''; 33 neighbours), have high PND, whereas words with few phonological neighbours, such as *en fangst* (''a catch''; 2 neighbours), have low PND.

PND has a positive effect on lexical development; early words have many phonological neighbours (Stokes, 2010; Storkel, 2004, 2009). PND shows differential effects in speech recognition and speech production. Experimental studies of speech recognition have found shorter reaction times for words with low PND than for words with high PND (Luce & Pisoni, 1998). The explanation for this is that words with many phonological neighbours activate more potential target words in the word recognition process than words with fewer phonological neighbours. Having many words activated slows down the process of selecting the correct target, and thus leads to longer reaction times in the experimental context (Luce & Pisoni, 1998). The opposite effect is found in speech production, where words with many phonological neighbours are more rapidly and more accurately activated than words with fewer phonological neighbours (Vitevitch, 2002). These differential effects have also been found in studies with language impaired speakers (Janse, 2009; Middleton & Schwartz, 2010).

Data on the PND of words in Norwegian was developed as part of an MA-study on the effects of imageability and PND on word perception and production (Ribu, 2012). Even though Norwegian has a fairly close orthography-to-phonology mapping, it was clear that a thorough calculation of PND had to be based on a substantial set of phonologically transcribed words. In order to accomplish this, a neighbourhood generator similar to the Language Independent Neighbourhood Generator of the University of Alberta (LINGUA)[5] was developed by the Text Laboratory at the University of Oslo. However, rather than generating orthographic neighbours, as LINGUA does, the generator developed for our purpose generated phonological neighbours for all the 1651 words in *Norwegian Words* based on the Bokmål section of *NorKompLeks* (Nordgård, 1998; cf. above).

In *Norwegian Words*, three categories of PND are distinguished: words with few neighbours (i.e. 0–2 neighbours), words with some neighbours (i.e. 3–18 neighbours) and words with many neighbours (19 or more neighbours). This categorisation is based on the same pattern as the categorisation of high, medium and low-frequency words. The 25% of the words with the highest number of phonological neighbours are classified as having many neighbours, the 25% of the words with the lowest number of phonological neighbours are classified as having few neighbours, and the middle 50% are classified as having some phonological neighbours. Again, for each word in the database information on the actual number of phonological neighbours is given, allowing for a recategorisation. For each word, one also finds a list of the phonological neighbours of that word, allowing for further exploitation of this feature in research and clinical practice.

[5]The software LINGUA is freely available online from: http://www.psych.ualberta.ca/~westburylab/downloads/lingua.download.html.

Imageability

Imageability is defined as the relative ease with which a word gives rise to a mental image or a sensory experience (Paivio, Yuille, & Madigan, 1968), and its effect on lexical processing has been explored in a number of studies, for instance Bird, Howard and Franklin (2003) and Prado and Ullman (2009). Furthermore, several studies have reported imageability to facilitate word learning (Ma, Golinkoff, Hirsh-Pasek, McDonough, & Tardif, 2009; McDonough, Song, Hirsh-Pasek, Golinkoff, & Lannon, 2011; Tardif, 1996, 2006).

Imageability data have been collected for varying numbers of words in several languages, including English (Bird, Franklin, & Howard, 2001; Paivio et al., 1968; Stadthagen-Gonzalez & Davis, 2006), Chinese (Ma et al., 2009), French (Desrochers & Thompson, 2009), Italian (Della Rosa, Catricalà, Vigliocco, & Cappa, 2010) and Japanese (Nishimoto, Ueda, Miyawaki, Une, & Takahashi, 2012). In most cases, imageability data have been gathered for nouns and verbs only, although there are exceptions, such as Bird et al. (2001) who included also adjectives and function words. A presentation of the collection of imageability data for the words in *Norwegian Words* and an analysis of these data with respect to interaction with word class, frequency and informant characteristics, can be found in Simonsen et al. (2013). For a more detailed account than the one we can give here, we refer to that article.

In order to establish imageability values for the words in *Norwegian Words*, a web-based survey was conducted during the spring of 2012. In total 399 informants were recruited from different organisations (hospitals, universities, large companies, etc.). Each informant was asked to rate only 100 words. When an informant agreed to participate by registering on a webpage, an SQL query randomly selected 100 of the words that had so far been given to the least number of informants. Thus, the word sets presented to each informant were random, but the ratings were distributed evenly across all the words. The information and instructions presented to the informants were based on Paivio et al. (1968). The informants were asked to rate the words (presented one at a time) on a 7-point scale (1 = ''no image'', 7 = ''clear image''). The informants could also report on a word as unknown or ambiguous, in which cases no rating of imageability could be made. The imageability value of a single word is the mean rating of that word in the survey. The mean number of imageability ratings for each word in *Norwegian Words* is 23.5 (SD 2.7).

In *Norwegian Words*, three levels of imageability are distinguished: low, medium and high. The categorisation is based on the same pattern as the categorisation of various degrees of frequency described above (25% in the low category, 50% in the medium category and 25% in the high category). Thus, words with imageability values ranging from 1 to 4.5 are classified as low imageability words, word with values ranging from 4.6 to 6.5 belong to the medium category, and words with ratings from 6.6 to 7 are classified as high imageability words. Here too, for each word in the database, information on the actual imageability value is given, allowing for different categorisations for different purposes.

Subjective age of acquisition

Subjective age of acquisition (AoA) is a measure of how early or late words are generally acquired. The AoA for a single word is calculated as the mean of the estimates made by a number of individual speakers regarding the age at which they know or think they acquired the relevant word. AoA has been shown to facilitate word processing in children as well as in adults. As a measure in lexical decision tasks it has been shown to be an even better predictor than other word property measures such as frequency and phonological neighbourhood density (Brysbaert & Cortese, 2011; Garlock, Walley, & Metsala, 2001).

To obtain AoA values for the words in *Norwegian Words* a web-based survey was conducted among adult, native speakers of Norwegian in the spring of 2012.[6] The informants were recruited from the same sources as for the collection of imageability ratings, and in total about 300 speakers responded. In the survey they were presented with one word at a time and asked to rate at which age they had acquired the word. For the rating, they had to use a scale ranging from 0 to 18, where 0 equals before one year of age, and 18 equals 18 years of age and above. If they could not remember when they had acquired a word (which we deem highly likely for many of the words, particularly those acquired early), they were explicitly asked to make a guess. Each speaker rated 120 words from among the 1651 words in the database, and the words were selected randomly for each speaker, but in such a way that all the words in the database received an equal number of ratings (cf. the description above for the collection of imageability ratings). On average, each of the mentioned ''best words'' from the cross-linguistic lexical task list was rated by 55.5 native speakers (SD 2.5); the rest of the words in *Norwegian Words* were on average rated by 12.9 native speakers (SD 1.5).

In the database, three categories of AoA are distinguished: words that are rated as acquired early (i.e. before age 3;7), words rated as acquired late (i.e. acquired after age 7;6) and words with a medium AoA (those rated as acquired between the ages 3;7 and 7;6). The categorisation is based on the same pattern as the categorisation of various degrees of frequency and imageability (25% in the early category, 50% in the middle category and 25% in the late category). Here too, for each word in the database information on the actual AoA value is given, allowing for a recategorisation.

Asking individuals to estimate at what age they acquired a certain word may seem a very dubious method for obtaining reliable data on age of acquisition. As mentioned, the informants were also explicitly asked to make a guess rather than refrain from responding if they could not remember. Łuniewska et al. (in progress) have investigated the reliability of this procedure by comparing AoA ratings for the mentioned 300 ''best words'' from 25 languages belonging to five different language families. They report significant and high correlations (ranging from 0.48 to 0.89) between all the 25 languages. Even though there are cross-linguistic differences as to when each word is estimated to be acquired, the order of acquisition of the words is very similar across the languages.

To measure the validity of the Norwegian subjective AoA ratings collected for *Norwegian Words*, we compared them to data from the Norwegian CDI study, consisting of parental reports on the lexical development of approximately 6500 children aged 0;8–3;0 (Kristoffersen & Simonsen, 2012; Simonsen et al., 2014). For the 458 items that are common to *Norwegian Words* and the Norwegian CDI vocabulary checklist, Hansen (in progress) has calculated a CDI-based AoA for each word, defined as the first age in months when at least half of the children in the sample are reported by their parents to produce it (Goodman et al., 2008; Ma et al., 2009; McDonough et al., 2011). Sixteen of the words do not reach the 50% limit by 36 months; all of which have a subjective AoA of 4;6 or more. The remaining 442 words are plotted in Figure 1, with their subjective AoA on the vertical axis and their CDI-based AoA on the horizontal axis.

Here, we see that the words systematically have a lower AoA according to the CDI data than according to our survey. Adults estimate that by age 3, they only knew 122 of the 442 items acquired by that age according to the CDI data. This means that either parents over-report on their children's lexical development, or adults underestimate their own vocabulary in retrospect. The latter seems most likely, as the Norwegian CDI results have been found to be valid and reliable (Kristoffersen et al., 2013; Simonsen et al., 2014).

[6]For the survey on AoA as well as for the survey on imageability, information was sent to The Norwegian Social Science Data Services, and the relevant permits were obtained.

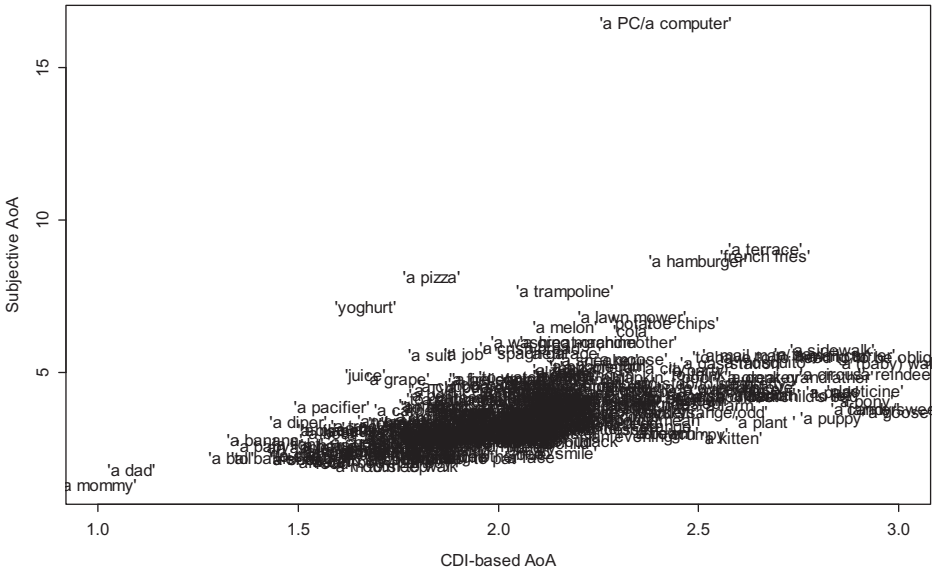

Figure 1. Correlation between subjective and CDI-based AoA for 442 items from *Norwegian Words*, labelled with English translations.

There are 12 words with a CDI-based AoA of age 3 or less, but a subjective AoA above age 6. These can be seen as outliers in Figure 1: *yoghurt, pizza, melon, trampoline, PC, cola, potetgull* ''potato chips'', *gressklipper* ''lawn mower'', *hamburger, pommes frites, terrasse* ''terrace'' and *tørketrommel* ''tumble dryer''. Apart from terrasse ''terrace'', all these words denote inventions and food products that are fairly new in Norway, and the results may be explained by the difference in birth date for the informants in the two groups we are correlating. The respondents in our subjective AoA study were all adult speakers (mean age: 38;5 years, SD: 15;4). The youngest of these speakers were born in 1995, whereas the oldest were born in 1939. The children, whose languages were reported on in the CDI study, were born in the period 2005–2010. Naturally, speakers born after the millennium will have acquired words like ''pizza'' and ''tumble dryer'' much earlier than speakers who grew up in the 1970s. There is a significant correlation between the two data sets ($r = 0.65$, $p < 0.001$ without the 11 inventions and food products). Similar results with a high correlation between subjective and objective age of acquisition measures are reported by Morrison, Chappell, & Ellis (1997).

Search options in the database

Norwegian Words allows for three types of searches. First, one can search for single words. By typing at least two letters in the search box, all words in the database (Norwegian words and English translations) which contain the given letter combination will automatically come up. One can then select one of these words and press a ''search'' button.[7] The search will provide all the available information about the given word in the database. For instance, a search for the word *ekkel* ''nasty'' will result in the information given in Table 2.

[7]It is unfortunately not yet possible to search for several words simultaneously (as a list).

Table 2. Result of a single word search in *Norwegian Words*.

Word	*Ekkel*
English translation	''Nasty''
Word class	Adjective
Imageability	Medium (4.52)
Subjective AoA	Medium (5.86)
Usage frequency	Medium (8619)
No. of phonological neighbours	Few (2)
Phonological neighbours	*Enkel, ekkelt*
Word length	
No. of letters	5
No. of sounds	4
No. of syllables	2
Word structure	
Derivation?	False
Compound?	False
Sound structure	
First sound fricative?	False
Starts in consonant cluster?	False

The second type of search that can be performed in *Norwegian Words* is to search for words that match with a selected set of properties and property values. One may for instance be interested in getting a list of disyllabic verbs with high imageability scores. By ticking the chosen values (verbs, high imageability and two syllables) in the relevant property boxes (word class, imageability and word length) and submitting the search, the webpage will return with the following list of words: *å bade* ''to bathe'', *å danse* ''to dance'', *å drikke* ''to drink'', *å dusje* ''to shower'', *å føde* ''to give birth'', *å male* ''to paint'', *å rake* ''to rake'', *å sitte* ''to sit'', *å stupe* ''to dive'' and *å tisse* ''to pee'', with information about word class, imageability value and number of syllables for each word.[8] The list is returned in a table format which may be downloaded as an Excel file.

In addition to selecting specific property values, one also has the opportunity to tick one or more property boxes in the right hand column on the webpage to get additional information about the words. One may for instance be interested in information about the frequency of these disyllabic, highly imageable verbs. This can be obtained by ticking ''Usage frequency'' in the right hand column in the search page. The subsequent search will then give the information that three of the 10 verbs are highly frequent (*å drikke* ''to drink'', *å føde* ''to give birth'' and *å sitte* ''to sit''), one is low in frequency (*å rake* ''to rake''), and the rest are medium in frequency. As mentioned above, for all the words the actual frequency value is also given, allowing whoever uses the database to draw the lines between the categories differently.

The third type of search available in *Norwegian Words* is to search for words from specific assessment tools. As described above, most of the words in the database are selected from tests and assessment tools for language acquisition and language impairment. In the search option, ''Assessment tools'' one can restrict the search to words from each of these test and assessment tools: the *Past Tense Test* (Ragnarsdóttir et al., 1999), the Norwegian versions of PALPA (Kay et al., 2009), CDI (Kristoffersen & Simonsen, 2012), CLT (Haman et al., in press) and VAST (Bastiaanse et al., 2006). The right hand column on the search page lists properties available in *Norwegian Words*, and one or more of these needs to be chosen to perform the search.

[8]These 10 verbs are of course not all the disyllabic, high imageability verbs in Norwegian, but they are the only verbs in the database that fit with the chosen criteria.

Utility of the database

Norwegian Words was deliberately designed to be useful for clinicians as well as researchers. Although the interests and needs of these groups may often be overlapping, they are not necessarily identical. Furthermore, at least in Norway, the various professions working with and doing research on language acquisition and language use in clinical and non-clinical populations groups (e.g. speech and language therapists, linguists, psychologists and kindergarten and primary school teachers), have very different backgrounds in general linguistics. To make the database accessible and useful to as many user groups as possible, we therefore decided to use everyday terminology whenever possible and to add short, explanatory texts to each of the properties in the database. These texts pop up when clicking/touching a small question mark in the search box for each property. Before launching the database, we also had a user evaluation of the interface by representatives of the various potential user groups (3 SLTs and 3 BA students of linguistics), and some minor amendments were done on the advice of this group.

In the first 18 months that the database existed – most of the time only in Norwegian – it was used by different people for a number of different purposes. In addition, we had inquiries from researchers internationally who wanted to use it, hence, the development of the English interface. Here, we briefly present some examples of what the database has been used for so far.

Facilitation of research is a main aim of *Norwegian Words*. Even though the database has not been publicly available for very long, it has already been used for research, for instance by Simonsen et al. (2013) in a study on the interaction of imageability, frequency and word class in different groups of speakers. It has also been used in several MA- and PhD-projects, including Ribu (2012) and Hansen (in progress). Ribu (2012) is a study of the impact of imageability and PND on perception and production of single nouns in three speakers with aphasia compared to a group of neurologically healthy adult speakers. Hansen (in progress) is a study of Norwegian children's acquisition of words as reported in the Norwegian CDI study (Kristoffersen & Simonsen, 2012; Simonsen et al., 2014) and four factors that have been shown to have an effect on word learning: word class, imageability, frequency and phonological neighbourhood density.

A second main aim of *Norwegian Words* is clinical utility. For instance, the database can be a useful tool which rapidly enables SLTs to select words for individually targeted assessment and/or treatment tasks for their clients. We have had positive feedback from SLTs who have used the database in this way in their clinical practice. One example refers to assessment of an adult with agraphia. In this case, *Norwegian Words* was used to find long words (3–6 syllables) starting with a consonant cluster, and these words were used, alongside other types of tasks, in the assessment phase to establish the client's level of difficulty with writing. A further example of clinical utility of the database is a current project on translation and adaptation of the *Comprehensive Aphasia Test* (Howard, Swinburn, & Porter, 2004) to a number of languages in Europe, including Norwegian. This project is carried out within the framework of COST Action IS1208, Collaboration of Aphasia Trialists (2013–2017). In the group working on the Norwegian adaptation, we have found the database to be a very helpful tool in searching for appropriate test items.

In addition to research, clinical practice and development of assessment tools, *Norwegian Words* has proven to be highly useful in teaching. We have used it for teaching purposes in an undergraduate course in psycholinguistics, both to illustrate points in lectures and seminars and as a part of student assignments. For example, the students were asked to analyse and discuss the word finding difficulties of a patient with anomia as manifested in a picture description task, and they were explicitly asked to use the database to inform their analysis.

These illustrations of how *Norwegian Words* has been used and is being used in research, clinical practice and teaching, are a few selected examples of the opportunities available. With the

addition of the English interface, we also anticipate added opportunities for bilingual and cross-linguistic applications.

Concluding remarks

There are some obvious limitations to a database such as *Norwegian Words*, the primary one being the size of the database. The more words that this kind of database contains, the better. However, the collection of data for various properties and the coding of the words can be costly. *Norwegian Words* was created without external funding, although we had some financial contributions from our university department to engage student assistants and pay for technical support. The number of words in the database is admittedly limited. However, each word is coded for a substantial number of properties.

Provided sufficient resources (time and money) there are of course opportunities for further development of the database, for instance adding new words with all or some of the already existing properties, and/or adding new properties to the already existing words. Currently, the concrete plans for further development of *Norwegian Words* include adding information about word frequencies in child directed speech (CDS) to the words in the database. As mentioned above, adult frequencies correlate with the acquisition of common nouns (Goodman et al., 2008). However, when CDS frequencies are applied, Goodman et al. (2008) report significant correlations between frequency and CDI-based AoA within all six broad lexical categories (common nouns, verbs, adjectives, people, closed class words and other). Hence, CDS frequencies for Norwegian might be a better starting point for research on language acquisition as well as development of assessment tools and tasks for clinical practice. We are currently gathering all the available semi-spontaneous Norwegian orthographically transcribed child – adult interaction in order to produce a frequency list for Norwegian CDS, and will subsequently add frequencies from this list to *Norwegian Words*.

As described above, the words in the database are gathered from various assessment tools for language acquisition and language disorders, on the assumption that this would enhance the clinical relevance of the database. While this may be a perfectly valid assumption, one could also argue that this may have led to a somewhat skewed selection of words, and that this could possibly affect the variation of the values in the database, at least for some properties. For instance, since we have only content words in our database, we have very few words that are extremely frequent in general in the language (such words would be function words). We also have very few words that are extremely infrequent in general. Examples of these are words that occur only once (hapax legomena) in a corpus. We also know that the mean imageability score in *Norwegian Words* is a bit higher than in most of the comparable studies (Simonsen et al., 2013), and we believe this reflects our choice of words (cf. also Nishimoto et al. (2012) for a similar finding).

As pointed out above, our division of words into categories such as high, mid and low, is based on an arbitrary, yet principled, choice. For all the relevant properties (i.e. phonological neighbourhood density, frequency, imageability and age of acquisition), the actual values are given for each word in the database, allowing for recategorisation of one or more of the properties by individual users of the database. In view of a potential expansion of the database to include more words, care should be taken, though, to include words from other domains and genres than assessment tools as well as from other lexical categories than nouns, verbs and adjectives.

Even though word properties are in one sense universal, we need information on them from each language separately; hence, the need for databases such as *Norwegian Words*. We hope *Norwegian Words* will serve as an inspiration for others to develop similar kinds of tools in their own language. Languages differ along structural parameters, posing different kinds of challenges

in acquisition as well as in language impairment. Furthermore, cultures, which partly form and are formed by language and language use, differ. This linguistic and cultural variation must be taken into account, also when searching for general aspects of language processing across the lifespan.

Acknowledgements

The database *Norwegian Words* was presented at the ICPLA conference in Stockholm, 11–13 June, 2014. We thank the audience as well as the reviewers for helpful questions and comments. We also thank Ingeborg Ribu for comments on an earlier version of the manuscript.

Declaration of interest

The authors report no conflicts of interest.

References

Bastiaanse, R., Lind, M., Moen, I., & Simonsen, H.G. (2006). *Verb- og setningstesten (VOST)*. Oslo: Novus.

Bird, H., Franklin, S., & Howard, D. (2001). Age of acquisition and imageability ratings for a large set of words, including verbs and function words. *Behavior Research Methods, Instruments, & Computers*, *33*, 73–79.

Bird, H., Howard, S., & Franklin, D. (2003). Verbs and nouns: The importance of being imageable. *Journal of Neurolinguistics*, *16*, 113–149.

Brysbaert, M., & Cortese, M. J. (2011). Do the effects of subjective frequency and age of acquisition survive better frequency norms? *The Quarterly Journal of Experimental Psychology*, *64*, 545–559

Della Rosa, P., Catricalà, E., Vigliocco, G., & Cappa, S. F. (2010). Beyond the abstract–concrete dichotomy: Mode of acquisition, concreteness, imageability, familiarity, age of acquisition, context availability, and abstractness norms for a set of 417 Italian words. *Behavior Research Methods*, *42*, 1042–1048.

Desrochers, A., & Thompson, G. L. (2009). Subjective frequency and imageability ratings for 3,600 French nouns. *Behavior Research Methods*, *41*, 546–557.

Garlock, V. M., Walley, A. C., & Metsala, J. L. (2001). Age-of-acquisition, word frequency, and neighbourhood density effects on spoken word recognition by children and adults. *Journal of Memory and Language*, *45*, 468–492.

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language*, *35*, 515–531.

Grunwell, P. (1981). *The nature of phonological disability in children*. London: Academic Press.

Guevara, E. (2010). NoWaC: A large web-based corpus for Norwegian. *Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop*. Los Angeles, CA: Association for Computational Linguistics.

Haman E., Łuniewska, M., & Pomiechowska, B. (in press). Designing cross-linguistic lexical tasks (CLTs) for bilingual preschool children. In S. Armon-Lotem, J. de Jong, & N. Meir (Eds.), *Methods for assessing multilingual children: Disentangling bilingualism from language impairment*. Bristol: Multilingual Matters.

Hansen, P. (in progress). What makes a word easy to learn? The effects of word class, imageability, phonological neighbourhood density and token frequency on Norwegian children's acquisition of content words. Department of Linguistics and Scandinavian Studies, University of Oslo.

Howard, D., Swinburn, K., & Porter, G. (2004). *Comprehensive aphasia test*. Hove: Psychology Press.

Janse, E. (2009). Neighbourhood density effects in auditory non-word processing in aphasic listeners. *Clinical Linguistics & Phonetics*, *23*, 196–207.

Jescheniak, J. D., & Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 824–843.

Jonkers, R., & Bastiaanse, R. (2007). Action naming in anomic speakers: Effects of instrumentality and name relation. *Brain and Language*, *102*, 262–272.

Kambanaros, M. (2013). Does verb type affect action naming in specific language impairment (SLI)? Evidence from instrumentality and name relation. *Journal of Neurolinguistics*, *26*, 160–177.

Kay, J., Lesser, R., & Coltheart, M. (2009). *Psykolingvistisk kartlegging av språkprosessering hos afasirammede (PALPA)* (Bredtvet kompetansesenter, Logopedtjenesten – Helse Bergen, Statped Vest & Øverby kompetansesenter, translation and adaptation). Oslo: Novus (Original published 1992).

Kittredge, A. K., Dell, G. S., Verkuilen, J., & Schwartz, M. F. (2008). Where is the effect of frequency in word production? Insights from aphasic picture-naming errors. *Cognitive Neuropsychology*, *25*, 463–492.

Korpijaakko-Huuhka, A.-M., & Lind, M. (2012). The impact of aphasia on textual coherence: Evidence from two typologically different languages. *Journal of Interactional Research in Communication Disorders*, *3*, 47–70.

Kristoffersen, G. (2000). *The phonology of Norwegian*. Oxford: Oxford University Press.

Kristoffersen, K. E., & Simonsen, H. G. (2012). *Tidlig språkutvikling hos norske barn. MacArthur-Bates foreldrerapport for kommunikativ utvikling*. Oslo: Novus.

Kristoffersen, K. E., Simonsen, H. G., Bleses, D., Wehberg, S., Jørgensen, R. N., Eiesland, E.A., & Henriksen, L. Y. (2013). The use of the Internet in collecting CDI data – an example from Norway. *Journal of Child Language*, *40*, 567–585.

Lind, M., Kristoffersen, K. E., Moen, I., & Simonsen, H. G. (2009). Semi-spontaneous oral text production: Measurements in clinical practice. *Clinical Linguistics & Phonetics*, *23*, 872–886.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*, 1–36.

Łuniewska, M., Haman, E., Armon-Lotem, S., Etenkowski, B., Pomiechowska, B., Anđjelković, D., Ayiomamitou, I., Blom, E., Boerma, T., Cantú Sánchez, M., Chiat, S., Dabašinskienė, I., Ege, P., Ehret, I., Engel de Abreu, P., Gagarina, N., Gatt, D., Gavarró, A., Håkansson, G., Hickey, T., Jensen de López, K., Kalninytė, A., Kambanaros, M., Kapalková, S., Kunnari, S., Levorato, C., Marinis, T., Nenonen, O., Nic Fhlannchadha, S., O'Toole, C., Odenbach, N., Polišenská, K., Popović, M., Ringblom, N., Rinker, T., Roch, M., Savić, M., Slancová, D., Southwood, F., Kronqvist, B. S., Thordardottir, E., Tsimpli, I., & Ünal, Ö. (in progress). Ratings of age of acquisition of 299 words across 25 languages. Is there a cross-linguistic order of words? Unpublished manuscript.

Ma, W., Golinkoff, R. M., Hirsh-Pasek, K., McDonough, C., & Tardif, T. (2009). Imageability predicts the age of acquisition of verbs in Chinese children. *Journal of Child Language*, *36*, 405–423.

McDonough, C., Song, L., Hirsh-Pasek, K., Golinkoff, R. M., & Lannon, R. (2011). An image is worth a thousand words: Why nouns tend to dominate verbs in early word learning. *Developmental Science*, *14*, 181–189.

Middleton, E. L., & Schwartz, M. F. (2010). Density pervades: An analysis of phonological neighbourhood density effects in aphasic speakers with different types of naming impairment. *Cognitive Neuropsychology*, *27*, 401–427.

Morrison, C. M., Chappell, T. D., & Ellis, A. W. (1997). Age of acquisition norms for a large set of object names and their relation to adult estimates and other variables. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *50*, 528–559.

Nishimoto, T., Ueda, T., Miyawaki, K., Une, Y., & Takahashi, M. (2012). The role of imagery-related properties in picture naming: A newly standardized set of 360 pictures for Japanese. *Behavior Research Methods*, *44*, 934–945.

Nordgård, T. (1998). Norwegian Computational Lexicon (NorKompLeks). *Proceedings of the 11th Nordic Conference of Computational Linguistics NODALIDA 98*. CST, Copenhagen.

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology, Learning, Memory and Cognition*, *76*, 1–22.

Pavlenko, A. (2009). Conceptual representation in the bilingual lexicon and second language vocabulary learning. In A. Pavlenko (Ed.), *The bilingual mental lexicon. Interdisciplinary approaches* (pp. 125–160). Bristol: Multilingual Matters.

Prado, E. L., & Ullman, M. T. (2009). Can imageability help us draw the line between storage and composition? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 849–866

Ragnarsdóttir, H., Simonsen, H. G., & Plunkett, K. (1999). The acquisition of past tense morphology in Icelandic and Norwegian children: An experimental study. *Journal of Child Language*, *26*, 577–618.

Ribu, I. S. B. (2012). *An image is worth a thousand sounds? On imageability and phonological neighborhood density effects in speech processing*. MA thesis, Department of Linguistics and Scandinavian Studies, University of Oslo, Oslo, Norway.

Simonsen, H. G. (1990). *Barns fonologi: System og variasjon hos tre norske og ett samoisk barn*. Department of Linguistics and Philosophy, University of Oslo, Oslo, Norway.

Simonsen, H. G., Hansen, P., & Łuniewska, M. (2012). Cross-Linguistic Lexical Tasks: Norwegian version (CLT-NO). A part of LITMUS COST IS0804 Battery. Unpublished material.

Simonsen, H. G., Kristoffersen, K. E., Bleses, D., Wehberg, S., & Jørgensen, R. (2014). The Norwegian Communicative Development Inventories – Reliability, main developmental trends and gender differences. *First Language*, *34*, 3–23.

Simonsen, H. G., Lind, M., Hansen, P., Holm, E., & Mevik, B. H. (2013). Imageability of Norwegian nouns, verbs and adjectives in a cross-linguistic perspective. *Clinical Linguistics & Phonetics*, *27*, 435–446.

Stadthagen-Gonzalez, H., & Davis, C. J. (2006). The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods*, *38*, 598–605.

Stokes, S. F. (2010). Neighborhood density and word frequency predict vocabulary size in toddlers. *Journal of Speech, Language, and Hearing Research*, *53*, 670–683.

Storkel, H. L. (2004). Do children acquire dense neighborhoods? An investigation of similarity neighborhoods in lexical acquisition. *Applied Psycholinguistics*, *25*, 201–221.

Storkel, H. L. (2009). Developmental differences in the effects of phonological, lexical and semantic variables on word learning by infants. *Journal of Child Language*, *36*, 291–321.

Tardif, T. (1996). Nouns are not always learned before verbs: Evidence from Mandarin speakers' early vocabularies. *Developmental Psychology*, *32*, 492–504.

Tardif, T. (2006). But are they really verbs? Chinese words for action. In K. Hirsh-Pasek, & R. M. Golinkoff (Eds.), *Action meets word: How children learn verbs* (pp. 477–498). Oxford: Oxford University Press.

Vitevitch, M. S. (2002). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology, Learning, Memory and Cognition*, *28*, 735–747.

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, *40*, 374–408.

Wilson, M. D. (1988). The MRC psycholinguistic database: Machine readable dictionary, Version 2. *Behavioural Research Methods, Instruments and Computers*, *20*, 6–11.

Notice of Correction
Changes have been made to this article since its original online publication date of 14 January 2015.