A Sound Approach to Text Processing: Between Experiments and Experience

Laura Winther Balling
Copenhagen Business School

Abstract
The area of text processing is an interesting one both for psycholinguists attempting to understand how language works and for those who focus on making texts accessible. However, understanding a text is a complex process that involves several different aspects, of which I discuss three main ones: comprehension, processing speed and ease, and reader reception, along with ways to study these aspects based on the reader and the text. A main argument in this chapter is that experiments should attempt to take these multiple aspects into account, and I describe two approaches that I have used to do so, and discuss their pros and cons. Based on this, some avenues of further research are outlined.

1. Introduction

Since working with Ocke-Schwen Bohn as his PhD-student, I have shared with him a devotion to understanding language through experiments. Experiments help us understand language processing on many levels, including speech perception, word recognition, and sentence processing. However, when we approach the level of text processing, we may be reaching the limits of what experiments can do, or at least what experiments can do alone: strict and sometimes artificial experimental manipulations and standard behavioural measures like response times provide a mechanistic and therefore too limited view of text processing. My argument in this chapter is that a sound approach to text processing should attempt to take

Anne Mette Nyvad, Michaela Hejná, Anders Højen, Anna Bothe Jespersen & Mette Hjortshøj Sørensen (Eds.), ***A Sound Approach to Language Matters – In Honor of Ocke-Schwen Bohn*** (pp. 355-388). Dept. of English, School of Communication & Culture, Aarhus University.

into account both what experiments tell us and what readers experience, attempting to bridge the gap between the language processing that happens in the psycholinguistic lab and the language processing that happens "in the wild".

The study of language processing in the psycholinguistic lab has traditionally been based on closely matched items in factorial designs, for instance comparing two different types of dative constructions using the same lexical material (e.g. Balling & Kizach, 2017). Stimuli appear without context and are generally constructed by the researchers rather than sampled from actual language use. In addition, the tasks that participants perform in the lab are often substantially different from language processing in the wild, including tasks such as acceptability ratings, which require the inclusion of ungrammatical sentences that we generally would not encounter in real written discourse; self-paced reading, which only gives access to one word at a time; and lexical decision, where we read or listen to both real words and constructed nonsense words and decide which are which. There are many advantages to the experimental approach, particularly in the ability to isolate and understand a particular aspect of processing, but in the case of text processing, this isolation arguably comes with too severe drawbacks.

In the following, I will discuss the problem of the gap in the study of text processing between language processing in the lab and language processing in the wild and ways in which it may be, if not bridged, then at least taken into account. The point of departure is an account in section 2 of what I see as the primary reasons to study text processing, followed by an outline in section 3 of three main aspects of text processing and how these may be studied. Section 4 focuses on attempts to bridge the gap between the lab and the wild, including both naturalistic experiments and other possible avenues of research. I focus here on the processing of written text, but some of the issues also apply to the study of oral discourse processing.

2. Why study text processing?

There are at least two main reasons for studying text processing: One is the general linguistic one of wanting to know how language works, with the specific psycholinguistic focus on understanding the relevant cognitive mechanisms. The level of text processing is particularly interesting – and complicated – in drawing on (the combination of) many other levels of processing. In this sense, the study of text processing is an attempt to understand the puzzle of how humans manage to read several hundred words

a minute while having to perform a range of different and in themselves rather complicated processing tasks, including recognising letters and combining them to form words, accessing the lexical representations of each of those words in a vocabulary of up to 150,000 words (Harley, 2008: 7), integrating them in phrasal and syntactic structures, and processing the discourse relations between them. Here, the influential Lexical Quality Hypothesis (Perfetti, 2007; Perfetti & Hart, 2002) argues that the quality and accessibility of the lexical representations is particularly important, but the other levels of processing and the coordination between them should of course not be overlooked.

The second reason is more practically oriented, namely the aim to improve text comprehension. This is for instance expressed in the Plain Language movement (see for instance *Federal Plain Language Guidelines*, 2011 for the US; Kjærgaard, 2016 for a discussion of the movement in the Nordic countries), as well as writing guides (e.g. Jacobsen & Jørgensen, 1992; Rozakis, 2000; Sorenson, 2010; Williams, 2005) and language policies, of which for instance those of the Danish courts (Kjærgaard, 2011a, 2011b, 2012) and Denmark's taxation authority have been carefully studied (Kjærgaard, 2015). These publications are not concerned with text processing *per se*, but the aim to improve comprehension and make texts more accessible should involve (psycholinguistic) considerations of how texts are actually processed and comprehended when read by different users.

3. Three main aspects of text processing

A main challenge when studying text processing is that it covers many different facets, not all of which are directly measurable. An obvious main issue – and in some respects in fact *the* main issue – is comprehension in the sense of understanding the contents of the text. This is an everyday activity for most literate people, but it remains hard both to define and to measure.

With respect to *defining* comprehension, it is important to consider both the depth and breadth of comprehension. The balance between these depends crucially on the purpose of reading – is it to look for a detail, get an overall sense of the topic, experience a narrative, or something else entirely – and on the motivation for reading, which could be interest or obligation or somewhere in between. Alternatively, we may think about purpose and motivation in terms of the three different types of goals for the reading of a given text outlined by Graesser, Singer, and Trabasso (1994):

default which is the goal of constructing an adequate situation model (see more on this concept below) for the text and is default in the sense of being generally applicable to most, if not all types of text processing; genre-based goals which are constrained by the type or genre of the text in question; and idiosyncratic goals which come close to what I refer to as the motivation for reading. In connection with the purposes and motivations for text comprehension, learning is an obvious issue in many contexts, but again, this is a phenomenon that comes in many varieties.

The variety of purposes and motivations for reading, and their influence on the process, mean that we may have to accept that text comprehension cannot be defined in isolation from the purpose of and motivation for the reading activity. This in turn becomes a methodological challenge in the empirical study of text processing, in that we must somehow define the level of reading expected of our participants, either by explicitly describing this or through specific task demands.

When it comes to *measuring* comprehension, there are multiple options which again depend on or may define the purpose of reading. An obvious choice in both text and sentence processing experiments (Bråten & Anmarkrud, 2013; Cop, Drieghe, & Duyck, 2015; Pham & Sanchez, 2018; Veldre & Andrews, 2018, to name but a few, from different domains) is multiple-choice or other forced-choice questions; this is a relatively quick and straightforward approach but potentially measures only rather superficial comprehension and relatively passive knowledge. More in-depth processing may be indexed by asking open questions (Balling, 2018) or requiring readers to recall the contents of texts they have read, and then scoring that recall for how many and how important ideas are recalled (e.g. Spyridakis & Isakson, 1998). Apart from the obvious drawback of being a more time-consuming research process, scoring of the replies relies to some extent on interpretation, particularly when it comes to the distinction between important and less important ideas.

In addition to the more generic objections to multiple-choice comprehension assessment, there is also evidence that different measures of comprehension measure different aspects of comprehension or different aspects of the text structure: Kintsch & Yarborough (1982) found that readers who had encountered texts with "good", conventional rhetorical structure performed better on questions about topic and main ideas of that text than those who read texts with "bad", unconventional rhetorical structure, while cloze test performance remained the same irrespective of rhetorical structure. It seems that overall rhetorical structure supports the

macro-processing indexed by recall rather than micro-processing indexed by cloze test performance. A further, extreme example of the difference between micro- and macro-level processing is that of quoting the Quran in Arabic without (otherwise) knowing Arabic (Kintsch, 1998: chapter 9), where a certain micro-level of learning is "measured" by repetition, but certainly not the kind of macro-level comprehension and learning that we are usually interested in when studying text processing.

A second important aspect of text processing is the ease and speed of the processing, which is partially an index of ease of comprehension but also depends on the efficiency of decoding as well as the more general language skills of the participants[1], which in turn vary with their reading proficiency. Experimental methods, including both standard behavioural measures like word reading time in self-paced reading and more advanced measurements like eye movements, are ideal for measuring speed, but cannot in themselves help us distinguish between text comprehension and decoding processes. To draw that distinction, there are broadly speaking two approaches: One is to measure decoding skills through one or typically more auxiliary tasks (see for instance the broad range of tasks used by Kuperman & Van Dyke, 2011). The other option is to conduct experiments with groups of participants with presumably similar decoding skills which is the typical approach when we run experiments with college students (e.g. most of the studies referenced in this text). However, even in such relatively homogeneous samples, we may well see substantial variation in decoding and general language processing skills, and the generalisability to "reading" in the abstract – to the extent that such a thing even exists, as discussed above – becomes questionable. We should also note that, although there is a correlation between text processing skills and general language skills, there is a substantial group of readers that read texts more poorly than we would expect based on their general language skills, as indicated by word reading ability (Perfetti & Adlof, 2012).

A third aspect of text processing, which is usually overlooked in the literature that focuses on text and discourse processes, but emphasised when the focus is on Plain Language and related approaches, is the reader's reception of the text and their resulting image of the sender. The methods for studying this are generally decidedly not experimental, but include comparative text analysis (e.g. Kjærgaard, 2011b), qualitative

[1] Because my focus is on text processing, I take the liberty of conflating these two, potentially quite different issues of decoding and general language skills, though in other contexts, it may be highly relevant to distinguish between them.

interviews (e.g. Garwood, 2014), and questionnaires (e.g. Kjærgaard, 2015). A particularly interesting approach in this field is the use of think-aloud protocols, a method that has also been used as a quasi-experimental paradigm in cognitive psychology (see e.g. Ericsson & Simon, 1980), to study the reading and reception of texts qualitatively (Kjærgaard, Gravengaard, Dindler, & Hjuler, 2018; Schriver, 1991).

One way of conceptualising these three different aspects of text processing – comprehension, processing speed and reception – is in relation to the general model of discourse processing that originates with van Dijk & Kintsch (1983) with later developments by Kintsch (1998) and others. This model includes five levels: surface code, text base, situation model, genre and rhetorical structure, and pragmatic communication. The surface code is the explicit lexical and syntactic contents of the text, which feed into the text base which is the reader's representation of the core semantic units of the text. The third level is the situation model which is the reader's representation of both the explicit contents of the text and the inferences drawn based on the text and existing knowledge. The fourth and fifth levels, like the first level, are oriented more towards the text/discourse than towards the receiver, with the fourth level being genre and rhetorical structure of the text or discourse, at various levels of granularity, and the fifth level the pragmatic communication, i.e. the message that the sender is trying to convey with the text or discourse.

This is not the only possible model of discourse processing, but it is one that is relatively broadly accepted and which provides a useful framework for understanding the elements involved in discourse and text processing. It also offers meaningful explanations for the different ways text processing may be impeded or even break down (Graesser & Millis, 2011). In relation to the aspects described here, speed and efficiency relate mostly to the first two levels of the model, while reader reception concerns levels 4 and 5, with comprehension understood as making sense of the text drawing on all levels but with a focus on level 3.

While the preceding parts of this section have focused on the readers and the reading process, an obvious further issue to consider is properties of texts. Here, an extensive literature has attempted to formulate readability indices that can measure the difficulty of a text, i.e. measure how difficult a reader will find a text to comprehend based on properties of that text. In a Danish context, the most common index is LIX (Björnson, 1968, cited by Klare, 1984), while in the US the most commonly used indices seem to be Flesch Reading Ease scale and the Flesch-Kincaid grade level scale that

was derived from that (Flesch 1943, Kincaid et al. 1975, both cited by Bailin & Grafstein, 2016). Most readability indices rely on some combination of word length or frequency with sentence length; for an overview see Bailin & Grafstein (2016), who also discuss many potential criticisms of standard readability measures. The major issue is the reliance on word and sentence lengths, which are correlated with, respectively, vocabulary difficulty and syntactic complexity but not perfectly so. For instance, relatively long and low-frequent words that consist of multiple well-known morphemes are not necessarily difficult to read because of their length (Bailin & Grafstein, 2016), and may in fact be easier to understand than their length would predict, due to their morphological structure supporting recognition (Balling, 2008). Similarly, longer sentences with simpler structure tend to be easier to read than shorter sentences with more complex syntactic structures, but this is not reflected in simple readability measures. In addition, the formulaic nature of the readability formulas means that the word and sentence length measures as well as frequency are assumed to have straightforwardly linear incremental effects, which is not necessarily the case (Bailin & Grafstein, 2016).

Another major issue, for readability formulas and for text processing in general, is text coherence, i.e. the logical structure of the text, and the explicit cohesive devices used to mark coherence. These are not captured by traditional readability measures, but are likely to play a central role to making sense of texts. A more recent attempt at automated capture of text readability, Coh-Metrix (for a comprehensive overview, see McNamara, Graesser, McCarthy, & Cai, 2014), does, as the name suggests, focus to a large extent on coherence, including multiple measures of markers of coherence, such as causal and referential cohesion. This approach is more refined than classical readability formulas, and generally also predicts text difficulty better, to the extent that we can measure that. However, the Coh-Metrix approach also suffers from one of the same fundamental problems as the more traditional readability formulas, namely the assumption that readability can be measured through some mechanistic combination of formal properties of the text (Bailin & Grafstein, 2016). Coh-Metrix uses more and more fine-grained variables, but it remains an issue for discussion whether this class of approaches really capture what we want to capture, and whether it is meaningful to attempt to measure readability based on texts alone, to the exclusion of the text user.

4. Bridging the gap

4.1 Naturalistic experiments

One way to attempt to bridge the gap between the lab and the wild in text processing research is to use experiments that are more naturalistic than the classical experiments described in the introduction. One way to do so is working with eye-tracking rather than experiments whose key measurements are based on explicit responses, like grammaticality judgment and self-paced reading. This method has been used more for studies of word and sentence processing than for studies of text, but at least since Rayner, Chace, Slattery, & Ashby (2006) also to investigate text and discourse processing. Rayner and colleagues found more and slightly but significantly longer fixations for complex texts, indicating that eye movements can be used as measurements of global text difficulty and text comprehension.

However, the use of eye-tracking methodology does not in itself make an experiment naturalistic. It does probably makes the reading process more similar to real-life reading processes than classical experimental tasks, but further steps are needed. One of them is investigating texts that are sampled from actual language use rather than constructed by the experimenter. For instance, I used authentic, only slightly edited descriptive and expository texts to investigate the effect of writing advice – such as 'avoid passives' and 'avoid nominalisations' that tend to show up in writing guides and language policies – on reading comprehension, in L1 Danish (Balling, 2013a) and L2 English (Balling, 2018). These studies are in some ways experimental in the sense outlined in the introduction, primarily because the investigation is based on two groups of participants each reading a different version of the same (sentential or phrasal) constructions. These experiments are nonetheless more naturalistic, and hence presumably more ecologically valid, than traditional experiments because they are based on authentic texts with minor experimental manipulations.

This use of authentic texts relies on three key design and analysis decisions: firstly, the experiments used eye tracking of reading. Secondly, the design and analysis relied on a regression approach where a range of relevant variables could be statistically controlled in the statistical analysis; since many predictor variables – including the frequency, predictability and length of words and constructions – by definition cannot be controlled beforehand in authentic texts, the statistical control becomes an absolute necessity. While length and frequency are relatively standard measures, predictability is harder to work with, leading to the third key

design decision of controlling predictability through conditional trigram frequency (originally inspired by MacDonald & Shillcock, 2003). The basic logic of this approach is that we index the predictability of a target word by taking the joint frequency of the target word and the two words preceding it and dividing it by the joint frequency of the two preceding words. For instance, for the highly predictable target word 'fløde' (cream) in the phrase 'rødgrød med fløde' (roughly translated as jelly with cream, a Danish dessert whose name is famously hard for non-Danes to pronounce):

$$p(fløde) = \frac{freq('rødgrød\ med\ fløde')}{freq('rødgrød\ med')}\text{, or } p(cream) = \frac{freq('jelly\ with\ cream')}{freq('jelly\ with')},$$

In other words, how often out of the times we find jelly with X is that X actually cream. In this case quite frequently, but of course the measure may also be used for very low predictabilities, and crucially also to gauge the differences between different low predictabilities by using tools from natural language processing (particularly the modified Kneser-Ney smoothing of Chen & Goodman, 1998) to deal with non-attested word bigrams and trigrams. This is in contrast to the standard method of measuring predictability, namely asking a group of participants to fill in cloze tests for the target words. The cloze method tends to assign the same zero probability to many words which are associated with probabilities which are different but not high enough for the word to show up in a cloze test (Yan, Kuperberg, & Jaeger, 2017). This lack of sensitivity at the low end of the scale is particularly problematic in view of the evidence that predictability effects are logarithmic in nature (Smith & Levy, 2013). The word trigram-based method described here has the additional advantage over cloze testing with human participants that, once the language model is trained, the extraction of the predictability measure for the relevant words is extremely fast. In the text processing experiments of (Balling, 2013a, 2018), the trigram-based predictability measure was averaged across the target constructions to index the average predictability of the words in the constructions.

These three design features – the use of eye-tracking, statistical control in regression analyses, and trigram models to index predictability – were used to allow the comparison of different types of target constructions in authentic descriptive and expository texts. The texts were only slightly edited to vary the versions of the target constructions between those forms that are recommended by writing guides and those that are labelled as

problem constructions, for instance actives vs. passives and sentential vs. nominal constructions (see an overview of the most prominent construction types in table 1). The original study by Balling (2013) showed no difference in fixation time between the recommended and problem constructions for highly skilled L1 readers of Danish. Balling (2018) tested a similar group of readers in their L2 English, investigating parallel differences for a lower-proficiency language but for readers with presumably similar decoding and general language skills. Again, this study did not show an effect on the fixation time on the different types of constructions. As a further attempt to encourage naturalistic but still somewhat controlled reading, the 2018 study used a set hypothetical but realistic comprehension frame for the texts and open questions to measure comprehension.

Problem	**Recommendation**	**Example**
Nominalisation	Verbal construction	- *is in relation to* + *relates to*
Reduced relative clause	Full relative clause	- *information contained* + *information that is contained*
Passive verb	Active verb	- *amounts covered* + *amounts we cover*
Long or complex words or sentences	Shorter sentences or words	- *be different* + *differ*

Table 1. Examples of the construction types investigated in Balling (2013a) and (2018), adapted from Balling (2018, table 1)

There are various possible reasons for this failure to detect an effect, aside from the substantive interpretation that the differences investigated do not in themselves matter, on which more below. These possible reasons fall into two groups: specific problems with these specific studies, and more general issues with this type of naturalistic experiment. Among the specific reasons is the obvious one that the power of the experiments may not have been sufficient to detect effects of this manipulation; the fact that the experiments did show effects of other predictors like construction length and the position of the construction in the sentence makes this explanation less likely, although it does remain a possibility. Turning to the design characteristics of the experiments, another possibility is that the texts were of too high quality (the manipulations on purpose did not disrupt the coherence of the texts), that the readers were too proficient to be affected

by these relatively minor manipulations, and, related to both these points, the possibility that the relevant manipulations – such as active vs. passive – pertain so exclusively to the surface code of the text that they do not affect processing in any measurable way, not even the relatively mechanistic reading time measures employed in these studies.

There are also more general potential problems with the two studies that arise because of the attempt to make the experiments as naturalistic as possible. One issue is the averaged conditional trigram probability used to index predictability: while this index works well as a predictor of the predictability of single words (Balling, 2013b), the averaged measure used in these two experiments and elsewhere (Balling & Kizach, 2017) may not be sensitive enough and is often only borderline significant. Another issue is that because of the use of authentic texts and a naturalistic set-up, the data are potentially quite noisy, particularly those from the 2018 study where participants were presented with full pages of text, while the 2013 study used sentence by sentence presentation which gives cleaner, or at least more "cleanable" eye-tracking data, but again also less natural reading.

4.2 Going more experimental: manipulating voice and givenness

Although we must always be careful with interpreting null results like the ones discussed above, it is conceivable that the construction type differences in themselves do not actually make a difference to text processing and comprehension. Nevertheless, it also remains a possibility that differences such as the one between active and passive constructions do in fact matter, but only when considered in conjunction with the key factor of text coherence. This possibility was investigated in a more strictly controlled experiment where the use of active vs. passive voice was manipulated in conjunction with the givenness of the agent and theme roles (Balling, in preparation).

The experiment used sentence-by-sentence self-paced reading of short constructed texts that were partly based on authentic texts from news outlets. Each text consisted of six pairs of sentences: target sentences with transitive main verbs in either the active or the passive voice and, immediately preceding each target sentence, a context sentence which set up either the agent or the theme of the target sentence as explicitly given, see table 2 for an example quadruple of related sentences. The dependent variable was reading time on the target sentences. In addition to the main 2*2 manipulation of voice and givenness, the analysis also included control

variables such as sentence length in characters, trial number (arguably indexing structural priming because the structures of target sentences were quite similar), and reading time on the immediately preceding context sentence.

CONTEXT SENTENCES		TARGET SENTENCES	
Agent of target sentence given	Another focus area is [DNA-investigations]$_{agent}$	Active	In the individual herd, [determine]$_{verb\ active}$ [DNA-investigations]$_{agent}$ [family relations between the giraffes]$_{theme}$
Theme of target sentence given	Another focus area is [family relations between the giraffes]$_{theme}$	Passive	In the individual herd, [determine]$_{verb\ passive}$ [family relations between the giraffes]$_{theme}$ by [DNA-investigations]$_{agent}$

Table 2. An example of a target sentence in active and passive voice versions, with context sentences setting the agent or the theme up as given. Each of the target sentences occurred with each of the context sentences, in different versions of the texts. Translated from the original Danish (preserving Danish V2 word order).

The underlying assumption of the manipulation is that a target sentence is easier to read if it is more coherent with the immediately preceding context sentence, and that such coherence may be at least partially achieved if the subject of the target sentence, which occurs as the first NP associated with the target verb, is explicitly given by the context sentence. This leads to the hypothesis that active sentences will be easier to understand if the agent – which takes the subject position in active sentences – is given, while passive sentences are easier to read if the theme – which is the subject of passive sentences – is given by the context sentence. However, this was not the case: a mixed-effects regression model (Bates, Mächler, Bolker, & Walker, 2015; Kuznetsova, Bruun Brockhoff, & Haubo Bojesen Christensen, 2016) in the R statistical environment (R Core Team, 2016) showed no significant interaction between voice and which thematic role was given by the context, no difference between active and passive sentences, and a main effect advantage for sentences in which the theme rather than the agent was given, an effect which is not in itself really interpretable.

This experiment was an attempt to further investigate the absence of an effect in the eye-tracking experiments described in section 4.1. At the same time, it was also an additional exploration of the continuum between experiments and experience, attempting to address two of the problems

with the previous studies – the predictability measurement problem and the noisiness of the data – that arose because of their naturalistic approach. The predictability issue was at least partly addressed by the systematic manipulation of voice and givenness on the same lexical material, and indeed an aggregated conditional trigram probability measure was not significant in these analyses. The same systematic manipulation could also contribute to less noisy data, compared to the many different types of constructions and the variations of them used in the experiments reported in section 4.1. However, this systematicity came at the cost of naturalness, with the constructed texts arguably coming across as too artificial to the readers. Finally, the reading time in the sentence-by-sentence self-paced task, which was partly chosen with the practical objective of being able to run multiple experiments simultaneously and thus get more participants, may be too insensitive to the after all relatively minor manipulation. Although the explicit givenness of first NP associated with the target verb does probably improve coherence, the difference between the two NPs was in practice one of relative givenness: in order to get the texts to work as texts, the not explicitly given NP was in many cases implicitly given.

4.3 Other avenues of research

While the approaches described in sections 4.1 and 4.2 should not be entirely discounted, the problems with them are also such that other avenues of research should be explored. There are several interesting possible perspectives, but common to them is the need to take into account multiple aspects of text processing.

One interesting way of doing this, which stays squarely in the experimental camp, is the approach of Kuperman, Matsuki, & Van Dyke (2018) who investigate the effects of the readers' cognitive and linguistic ability, the linguistic properties of the text, and the temporal dynamic of the reading process, and crucially also the interaction and relative weights of these. This goes some way towards understanding the many levels of text processing and the emphasis on interactions is both novel and crucial. More generally, reader proficiency in a broad sense is an important aspect to take into account, and on trend in relation to the recent emphasis on individual differences in language processing (for an overview, see Kidd, Donnelly, & Christiansen, 2018). However, the clear experimental focus of Kuperman et al. (2018) still means that the more experience-related issues of in-depth comprehension and reader reception are not clearly addressed.

Given the interdependence of the different aspects of text processing outlined in this chapter, an ideal would be joint consideration of the multiple aspects – including comprehension, speed, and reception – while looking at both the narrowly processing-oriented aspects measured in experiments and the experience of language users in the wild. This ideal may be partially implemented by mixed-methods approaches, though it remains to be worked out exactly how to interpret potentially diverse results together. As a compromise, the reception aspect and variations in comprehension beyond what may be measured in multiple-choice questions should as a minimum be considered as valid concerns in relation to text processing; conversely, the more text processing oriented aspects should in turn be seriously evaluated rather than just assumed in writing guides and language policies.

6. References

Bailin, A., & Grafstein, A. (2016). *Readability: Text and Context*. Houndmills: Palgrave Macmillan.

Balling, L. W. (2008). *Morphological Effects in Danish Auditory Word Recognition*. PhD-thesis, University of Aarhus.

Balling, L. W. (2013a). Does Good Writing Mean Good Reading ? An Eye-tracking Investigation of the Effect of Writing. *Fachsprache*, *35*(1-2), 2-23.

Balling, L. W. (2013b). Reading authentic texts: What counts as cognate? *Bilingualism: Language and Cognition*, *16*(3), 637-653. https://doi.org/doi:10.1017/S1366728911000733

Balling, L. W. (2018). No Effect of Writing Advice on Reading Comprehension. *Journal of Technical Writing and Communication*, *48*(1), 104-122. https://doi.org/10.1177/0047281617696983

Balling, L. W., & Kizach, J. (2017). Effects of Surprisal and Locality on Danish Sentence Processing: An Eye-Tracking Investigation. *Journal of Psycholinguistic Research*, *46*(5), 1119-1136. https://doi.org/10.1007/s10936-017-9482-2

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software*, *67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Björnson, C. H. (1968). *Läsbarhet*. Stockholm: Liber.

Bråten, I., & Anmarkrud, Ø. (2013). Does naturally occurring comprehension strategies instruction make a difference when students read expository text? *Journal of Research in Reading*, *36*(1), 42-57. https://doi.org/10.1111/j.1467-9817.2011.01489.x

Chen, S. F., & Goodman, J. (1998). An empirical study of smoothing techniques for language modeling. *Harvard University Technical Report*, *TR-10-98*.

Cop, U., Drieghe, D., & Duyck, W. (2015). Eye movement patterns in natural reading: A comparison of monolingual and bilingual reading of a novel. *PLoS ONE*, *10*(8), 1-38. https://doi.org/10.1371/journal.pone.0134008

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*(3), 215-251. https://doi.org/http://dx.doi.org/10.1037/0033-295X.87.3.215

Federal Plain Language Guidelines. (2011). Retrieved from http://www.plainlanguage.gov/howto/guidelines/FederalPLGuidelines/FederalPLGuidelines.pdf. April 30, 2012.

Flesch, R. (1943). *Marks of Readable Style: A Study in Adult Education*. Teachers College, Columbia University.

Garwood, K. (2014). *Plain, but not Simple: Plain Language Research with Readers, Writers and Texts*.

Graesser, A. C., & Millis, K. (2011). Discourse and Cognition. In T. A. van Dijk (Ed.), *Discourse Studies: A Multidisciplinary Introduction* (pp. 126-142). London: SAGE Publications. https://doi.org/10.4135/9781446289068.n16

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, *101*(3), 371-395. https://doi.org/10.1037/0033-295X.101.3.371

Harley, T.A. (2008). *The Psychology of Language. From Data to Theory* (3rd ed.). Hove & New York: Psychology Press.

Jacobsen, H. G., & Jørgensen, P. S. (1992). *Håndbog i Nudansk* (2nd ed.). København: Politikens Forlag.

Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual Differences in Language Acquisition and Processing. *Trends in Cognitive Sciences*, *22*(2), 154-169. https://doi.org/10.1016/j.tics.2017.11.006

Kincaid, J.P., Fishburne, R.P., Rogers, R.L., & Chissom, B.S. (1975). *Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for Navy enlisted personnel*. Research Branch Report 8-75. Millington, TN, Naval Technical Training, U. S. Naval Air Station, Memphis, TN.

Kintsch, W. (1998). *Comprehension. A paradigm for cognition*. Cambridge, UK: Cambridge University Press.

Kintsch, W., & Yarborough, J. C. (1982). Role of rhetorical structure in text comprehension. *Journal of Educational Psychology*, *74*, 828-834.

Kjærgaard, A. (2011a). Det laaange seje træk, del 2. Mere om Sprogpolitik for Danmarks Domstole. *Nyt Fra Sprognævnet*, *2011/2*, 7-12.

Kjærgaard, A. (2011b). Nytter det? – Om de tekstlige effekter af sprogpolitiske projekter i offentlige institutioner. *Nydanske Sprogstudier*, *40*, 90-116.

Kjærgaard, A. (2012). Fra lidenskab til ligegyldighed – En caseanalyse fra Danmarks Domstole af et sprogpolitisk projekts (manglende) gennemslagskraft. *Sakprosa*, *4*(1), 1-28.

Kjærgaard, A. (2015). Påvirker omskrivninger af tekster fra det offentlige borgernes forståelse – og hvordan? *Sakprosa*, *7*(2), 1-25.

Kjærgaard, A. (2016). The organisation of the plain language movement in Denmark. In P. Nuolijärvi & G. Stickel (Eds.), *Language use in public administration. Theory and practice in the European states. Contributions to the EFNIL Conference 2015 in Helsinki.* (pp. 123-134). Helsinki: EFNIL.

Kjærgaard, A., Gravengaard, G., Dindler, C., & Hjuler, S. (2018). Tænke højt-protokoller. En metode til at undersøge modtageres tekstforståelse og -oplevelse. *Nydanske Sprogstudier*, *54*.

Klare, G. R. (1984). Readability. In P. D. Pearson (Ed.), *Handbook of Reading Research*, pp. 681-741. Mahwah, NJ: Lawrence Erlbaum.

Kuperman, V., Matsuki, K., & Van Dyke, J. A. (2018). Contributions of reader- and text-level characteristics to eye-movement patterns during passage reading. *Journal of Experimental Psychology: Learning, Memory & Cognition*, 44(11), 1687-1713. https://doi.org/10.1037/xlm0000547

Kuperman, V., & Van Dyke, J. A. (2011). Effects of individual differences in verbal skills on eye-movement patterns during sentence reading. *Journal of Memory and Language*, *65*(1), 42-73. https://doi.org/10.1016/j.jml.2011.03.002

Kuznetsova, A., Bruun Brockhoff, P., & Haubo Bojesen Christensen, R. (2016). lmerTest: Tests in Linear Mixed Effects Models. Retrieved from https://cran.r-project.org/package=lmerTest

MacDonald, S. A., & Shillcock, R. C. (2003). Low-level predictive inference in reading: the influence of transitional probabilities on eye movements. *Vision Research*, *43*, 1735-1751.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated Evaluation of Text and Discourse with Coh-Metrix*. New York: Cambridge University Press.

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, *11*(4), 357-383. https://doi.org/10.1080/10888430701530730

Perfetti, C. A., & Hart, L. (2002). The Lexical Quality Hypothesis. In L. Verhoeven, C. Elbro, & P. Reitsma (Eds.), *Precursors of functional literacy* (pp. 189-213).

Perfetti, C., & Adlof, S. M. (2012). Reading Comprehension : A Conceptual Framework from Word Meaning to Text Meaning. In J. Sabatini & E. Albro (Eds.), *Assessing reading in the 21th century: Aligning and applying advances in the reading and measurement sciences* (pp. 3-20).

Pham, H., & Sanchez, C. A. (2018). Text Segment Length Can Impact Emotional Reactions to Narrative Storytelling, to appear in *Discourse Processes*, *0*(0), 1-19. https://doi.org/10.1080/0163853X.2018.1426351

R Core Team. (2016). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from http://www.r-project.org/

Rayner, K., Chace, K. H., Slattery, T. J., & Ashby, J. (2006). Eye Movements as Reflections of Comprehension Processes in Reading. *Scientific Studies of Reading*, *10*, 241-255.

Rozakis, L. (2000). *Complete Idiot's Guide to Writing Well*. East Rutherford, NJ, USA: Penguin Putnam.

Schriver, K. A. (1991). Plain Language for Expert or Lay Audiences: Designing Text Using Protocol-Aided Revision, Technical Report No. 46, Center for the Study of Writing. Available from https://files.eric.ed.gov/fulltext/ED334583.pdf

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302-319. https://doi.org/10.1016/j.cognition.2013.02.013

Sorenson, S. (2010). *Webster's New World Student Writing Handbook* (5th ed.). Hoboken, New Jersey: Webster's New World.

Spyridakis, J. H., & Isakson, C. S. (1998). Nominalizations vs. denominalizations: do they influence what readers recall? *Journal of Technical Writing and Communication*, *28*, 163-188.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of Discourse Comprehension*. New York: Academic Press.

Veldre, A., & Andrews, S. (2018). Beyond cloze probability: Parafoveal processing of semantic and syntactic information during reading. *Journal of Memory and Language*, *100*, 1-17. https://doi.org/10.1016/j.jml.2017.12.002

Williams, J. M. (2005). *Style. Ten lessons in clarity and grace* (8th ed.). New York: Pearson Longman.

Yan, S., Kuperberg, G. R., & Jaeger, T. F. (2017). Prediction (Or Not) During Language Processing. A Commentary On Nieuwland et al. (2017) And Delong et al. (2005). *bioRxiv*. Retrieved from http://biorxiv.org/content/early/2017/05/30/143750.abstract.