Picture Description in the Assessment of Connected Speech Intelligibility in Parkinson’s Disease: A Pilot Study

Inga-Lena Johansson\textsuperscript{a}  Christina Samuelsson\textsuperscript{a, b}  Nicole Müller\textsuperscript{a, c}

\textsuperscript{a}Department of Biomedical and Clinical Sciences/Speech and Language Pathology, Linköping University, Linköping, Sweden; \textsuperscript{b}Department of Clinical Science, Intervention and Technology (CLINTEC), Karolinska Institute, Solna, Sweden; \textsuperscript{c}Department of Speech and Hearing Sciences, University College Cork, Cork, Ireland

Keywords
Dysarthria · Parkinson’s disease · Intelligibility · Assessment · Picture description

Abstract

Introduction: Assessment of intelligibility in dysarthria tends to rely on oral reading of sentences or words. However, self-generated utterances are closer to a client’s natural speech. This study investigated how transcription of utterances elicited by picture description can be used in the assessment of intelligibility in speakers with Parkinson’s disease. Methods: Speech samples from eleven speakers with Parkinson’s disease and six neurologically healthy persons were audio-recorded. Forty-two naïve listeners completed transcriptions of self-generated sentences from a picture description task and orally read sentences from the Swedish Test of Intelligibility, as well as scaled ratings of narrative speech samples. Results: Intelligibility was higher in orally read than self-generated sentences and higher for content words than for the whole sentence in self-generated sentences for most of the speakers, although these within-group differences were not statistically significant at the group level. Adding contextual leads for the listeners increased intelligibility in self-generated utterances significantly but with individual variation. Although correlations between the intelligibility measures were at least moderate or strong, there was a considerable inter- and intra-speaker variability in intelligibility scores between tasks for the speakers with Parkinson’s disease, indicating individual variation of factors that impact intelligibility. Intelligibility scores from neurologically healthy speakers were generally high across tasks with no significant differences between the conditions. Conclusion: Within-speaker variability supports literature recommendations to use multiple methods and tasks when assessing intelligibility. The inclusion of transcription of self-generated utterances elicited by picture description to the intelligibility assessment has the potential to provide additional information to assessment methods based on oral reading of pre-scripted sentences and to inform the planning of interventions.

Introduction

This study investigates transcription of utterances elicited by picture description in the assessment of intelligibility in speakers with Parkinson’s disease. In Parkinson’s disease, progressive hypokinetic dysarthria is common...
and can affect speech functions on many levels, such as breath support, voice, and articulation [1, 2]. Characteristics of hypokinetic dysarthria include monotonous speech, reduced stress, varying and sometimes accelerating speech rate, use of short phrases, and imprecise articulation [2]. A common and key problem in dysarthria, regardless of type, is decreased speech intelligibility [3–5], which can be defined as understandability of speech [6], the performance by a listener to recognize the words and/or the sounds produced [7], or the degree to which the speaker’s intended message is recovered by the listener [8]. In Parkinson’s disease, there is a known heterogeneity among speakers [9–11], and possible task-dependent performance effects [4, 12–16] which need to be considered in assessment for planning interventions to optimize communication. This includes assessment of intelligibility.

The Concept of Intelligibility

It is common to separate intelligibility of the acoustic speech signal from comprehensibility of the content the speaker is conveying, which can also include contextual information and body language [2, 3, 17–20]. In the literature, the former is also referred to as signal-dependent intelligibility, and the latter as signal-independent intelligibility [18, 19]. Measures of signal-dependent and signal-independent intelligibility give different but complementary information about dysarthric speech, and they require different assessment methods [17, 19]. However, the concepts are not mutually exclusive and there is a complex interaction between them [19]. We have therefore chosen to use the term intelligibility, since the primary focus is on intelligibility of the speech signal, although some aspects of the study also expand into an intermediate area towards signal-independent intelligibility or comprehensibility.

Assessment of Connected Speech Intelligibility

Yorkston et al. [6] state that few measures other than intelligibility of connected speech allow for an overall evaluation of the interaction of processes involved in dysarthric speech. To this could be added that measures of intelligibility should capture as closely as possible the characteristics of natural speech. Many intelligibility assessments include word recognition tests [19, 21]. However, speech production of isolated words is different from connected speech, and as a consequence, intelligibility scores from single word tests might not be representative of intelligibility in connected speech [22, 23].

Samples of connected speech are usually elicited through oral reading or repetition. These tasks permit control over linguistic variables and provide consistency across samples [6, 24], but they also differ from natural speech. In addition, performance effects have been noted in Parkinson’s disease with higher intelligibility scores in oral reading than in spontaneous or conversational speech [12–15]. Speech elicited through narration or conversation allows speakers to use their individual speech style and to adjust their utterances to their own speech abilities, which is probably employed in individually different ways and is likely to change over time in a progressive disease. Intelligibility tasks based on spontaneous or conversational speech would be preferable in terms of face validity but reduce the control over content and utterance structure compared to oral reading tasks [8]. Picture description allows for self-generated utterances and at the same time provides more control over content [24]. Assessment of intelligibility through scaled ratings of speech elicited by picture description has been used with children [25] and adult patients with oral cancer [26]. Picture description has also been used for elicitation of speech for other purposes, such as segmental and prosodic analysis of speech in children with motor speech disorders [27], and to linguistically compare oral and written discourse in patients with aphasia [28]. Given these previous applications to heterogeneous populations, picture description, therefore, is a promising technique to be added to more established assessment methods.

Listener Ratings of Intelligibility

Transcription or scaled ratings are common listener tasks for the measurement of connected speech intelligibility [19, 21, 24–26]. In the literature, transcription is commonly described as an objective measure of intelligibility, while scaling is considered more subjective [3, 19, 29, 30]. Orthographic transcription can be a more consistent measure than, for example, per cent estimates [31]. Scales for rating or estimation are simple to conduct and give insights into listeners’ perceptions of the speech and may more closely represent the functional level of communication; however, the validity, reliability, and reproducibility of scaling procedures have been questioned [3, 19, 21, 24, 31, 32]. For example, listeners may use different criteria when they are asked to do a scaled rating of how understandable they perceive a sample of speech to be [25, 33]. For ratings in the mid-range of a scale, as could be the case for speakers with moderately affected intelligibility, there seems to be a greater inter-rater variability than for ratings near the end points [33, 34].
Assessment of Intelligibility in Parkinson’s Disease

Contextual Cues

Syntactic and semantic cues can increase intelligibility in dysarthric speech [20, 35, 36]. Therefore, the use of sentences with low predictability is recommended for assessment [24]. The types of words in an utterance can also impact intelligibility. In a study on speakers with cerebral palsy, Hustad [37] found that non-information-bearing function words were transcribed more accurately than content words and discussed the importance of listener focus on information-bearing words, and that possible reasons for the noted difference could be the simpler structure and syntactic predictability of function words.

Adding contextual cues when assessing intelligibility can more closely resemble a conversational setting, which usually has some predictability [6]. Listeners’ topic knowledge has been found to benefit intelligibility, but with mixed results [6, 38]. In Utianski et al.’s [38] study, including speakers with ataxic and hypokinetic dysarthria, there was a significant effect of topic knowledge for speakers with ataxic dysarthria, but not for the participants with hypokinetic dysarthria. However, this study was based on sentence reading, and it is possible that performance effects, which have been noted in oral reading for speakers with Parkinson’s disease [12–15], might equalize the results from the different conditions.

Assessment of Intelligibility in Swedish

In the Swedish context, where the present study was conducted, the Swedish Test of Intelligibility (STI) [39] is the most widely used assessment material for intelligibility in dysarthria. It consists of both word recognition from a phonetically balanced word list by multiple choice among phonetically related alternatives (rhyme words), and transcription of sentences. The sentences in the STI are syntactically well formed nonsense sentences. They each consist of four to five content words (nouns, adjectives, and verbs), and some sentences also contain indefinite articles and/or a conjunction. The client reads the words and sentences aloud. The STI is a useful tool for reproducible measures of intelligibility of the acoustic speech signal per se. However, as discussed above, words and sentences elicited by oral reading may not necessarily be a good representation of a person’s natural speech. In addition, the nonsense sentences in the STI can challenge the speaker’s reading skills [23].

Aims of the Present Study

Intelligibility assessment would preferably be based on spontaneous or conversational speech in terms of face validity and ecological validity. Where spontaneous speech is assessed, its intelligibility is usually measured by scaled ratings. However, the validity and reliability of scaling procedures for measuring intelligibility have been questioned [3, 19, 21, 31]. Tasks based on oral reading permit control over linguistic variables and provide consistency across samples. However, oral reading differs from natural speech and previous studies of intelligibility in Parkinson’s disease have noted performance effects related to oral reading [12–15]. Currently, there is a lack of other methods for reproducible measurement of intelligibility in tasks that resemble natural speech as closely as possible. One specific challenge when using extemporaneous speech for assessment of intelligibility is to know what persons with dysarthria are saying. Since picture description provides more control over content, intelligibility assessment through transcription of utterances could present a valuable compromise between naturalness and control. Another advantage of picture description is that it easily can give the opportunity to assess the effect of context on intelligibility in self-generated utterances. To the best of our knowledge, there are no published studies investigating assessment of intelligibility in dysarthria by transcription of utterances elicited through picture description.

The first aim of the present study was to try out transcription of self-generated utterances elicited through picture description as a method for the assessment of intelligibility in speakers with Parkinson’s disease. Based on the literature review, we especially wanted to explore if adding transcription of self-generated utterances to the assessment has the potential to provide additional information about: (1) intelligibility in oral reading versus self-generated speech, (2) intelligibility of content words in self-generated sentences, and (3) effects of listener knowledge of context.

A second aim was to investigate how intelligibility scores from self-generated utterances correlate with scores based on sentences from the STI, elicited by oral reading, and scaled ratings. Possible effects of utterance length were also controlled for.

Materials and Methods

Participants

Persons experiencing speech problems due to Parkinson’s disease were recruited through a Parkinson’s disease association and through speech and language pathologists in a neurorehabilitation clinic. Eleven speakers with Parkinson’s disease participated. Six persons without neurological or other speech problems were also recruited to provide data from healthy speakers for comparison. Group characteristics are presented in Table 1.

Group characteristics are presented in Table 1.
All participants were native Swedish speakers living in the south-east of Sweden. The participants reported no previous history of speech or language disorders. Occurrence of hearing impairment and use of hearing aids was explored through self-report but was not an exclusion criterion for this study, since hearing loss is common among the elderly [40] and often co-occurs with Parkinson’s disease [41]. Overall disease severity was rated with the Hoehn and Yahr-scale [42], which has a range from stage 1 or very mild symptoms to stage 5 or severe symptoms with constant need of nursing care. None of the participants had a diagnosis of dementia.

For the intelligibility measures, we recruited 42 naive listeners with Swedish as the primary language and without self-reported hearing impairment or problems with reading or writing. We chose to use naive listeners because they might be more representative of the general public. The listeners were recruited in the same geographic area in the south-east of Sweden as the speakers. Twenty-eight of the listeners were SLP students at the beginning of their first semester and had not yet received any training in phonetics, 18–46 years (M = 25; SD = 7.44). Six of the listeners were recruited outside the university. Age distribution among the listeners or staff from other departments, and six listeners were linguistics, or speech pathology. Eight listeners were university students or staff from other departments, and six listeners were recruited outside the university. Age distribution among the listeners was 18–46 years (M = 25; SD = 7.44). Six of the listeners were male and 36 were female.

Procedure

Collection of Speaker Data

The participants’ speech was audio-recorded either in university locations or in their homes, according to their choice. Seven of the participants with Parkinson’s disease opted for their homes and the rest of the participants for university locations. Regardless of location, the recording environment was quiet, without background noise from, for example, refrigerators or traffic, in a location where the participant could sit comfortably during recording at approximately 35-cm distance from a table microphone. The participants also chose their preferred time, based on their experiences of what time of day they were functioning at their best.

Each recording session lasted in total for about 50 min and included the following elements: (a) an interview about personal and disease background based on questionnaires sent out to the participants in advance, (b) the short form of the Swedish dysarthria assessment [39], (c) cognitive screening with the Mini-Mental State Examination – Swedish revision (MMSE-SR) [43], and (d) three connected speech tasks. These included (1) a picture description task from a Swedish neurolinguistic assessment [44], where the picture showed an outside restaurant scene with several actions in progress, and the participants were instructed to imagine that the listener was not familiar with and could not see the picture, (2) oral reading of 3 lists of 10-sentences each (total 30) from the STI State Examination – Swedish revision (MMSE-SR) [43], and (3) a free narrative, where the speakers were asked to talk about a topic of their own choice, such as a hobby or an event in their lives.

Recording equipment was a Marantz PMD 660 and a tabletop microphone AKG C 535 EB. The input recording level was adjusted for a level minimizing risk for sound clipping or noise in normal conversational loudness at 35-cm distance from the microphone. The chosen settings for the recordings were a sampling rate of 48 kHz, 16 bit uncompressed PCM/WAV-format.

Speech Materials for the Listener Tasks

Picture Description. Four utterances per speaker were derived from the picture description task, each consisting of a full declarative sentence. To enable exploration of intelligibility in content-bearing parts of the utterances, each sentence contained a content word not predictable from other words in the utterance. The non-predictability of target content words was tested and confirmed by a sentence-completion task administered to three SLPs. To ensure correct interpretation of what the speakers had said and consistent scoring of listener transcriptions, ambiguous utterances were judged and transcribed also by two SLPs who were not involved in the project. The inter-judge agreement was 100% between each of

<table>
<thead>
<tr>
<th>Table 1. Group characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Gender</td>
</tr>
<tr>
<td>Age, years</td>
</tr>
<tr>
<td>MMSE-scoreb</td>
</tr>
<tr>
<td>Hearing impairment/use of hearing aids</td>
</tr>
<tr>
<td>Time since diagnosis, years</td>
</tr>
<tr>
<td>Disease severity (H&amp;Y)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Dysarthria severityc</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

* Mann-Whitney U. b Maximum possible score = 30. c Assessed by 1st author with the short form of the Swedish Dysarthria Assessment (Hartelius [39]).
the independent judges, who judged and transcribed extracted utterances without contextual information, and the 1st author, who had carried out the recordings of the speech samples and therefore had knowledge of the content of the picture description utterances. The largest proportions of target content words were nouns and verbs for both the participants with Parkinson’s disease (nouns: 61%; verbs: 23%) and the neurologically healthy speakers (nouns: 62%, verbs: 29%), and the remaining minor proportion consisted of adjectives and adverbs. There were no significant differences in distribution across word classes for target content words between the groups.

To be able to control for effects of word length on intelligibility, the length of the target content word in three of the sentences was one, two, or three syllables, respectively. The fourth sentence contained the most articulatorily complex content word found in the remaining utterances fulfilling the criteria above. For some speakers, it was not possible to find four separate sentences fulfilling the criteria above and in these cases, two of the target content words were derived from the same sentence. Examples of sentences from picture descriptions are shown in the online supplementary material (see www.karger.com/doi/10.1159/000521906 for all online suppl. material).

**Sentences from the STI.** Four sentences per speaker from the recordings of STI-sentences were randomly chosen. Sentences, where words were replaced or omitted during the oral reading were excluded. Examples of STI-sentences are shown in the online supplementary material.

**Free Narratives.** An extract of connected speech of around 15 s (M = 16.62; SD = 1.91) was obtained from the free narrative task for each speaker. A criterion for the extract for the listener task was that the content should be coherent, without topic shifts. The length of the extract was based on analysis of the recorded speech material, where the longest coherent narrative from the speaker with the greatest speech production difficulties was 15 s. To keep conditions between speakers similar, the length of the extracts for the other speakers was adjusted to the point closest to 15 s, where the criterion of coherent content was met. Examples of free narrative extracts are shown in the online supplementary material.

**Preparation of Listener Sets**

To avoid confounds from listener familiarization and fatigue, the speech material was divided into six listener sets. Figure 1 shows an overview of how the listener sets were organized.

In each subset of sentences from picture description, the sentences from the eleven speakers with Parkinson’s disease and the six neurologically healthy speakers were proportionally distributed (65%/35%), as were the extracts from a narrative speech in each listener set. The ordering of sentences or extracts of narrative speech within each subset was randomized. The sentences from picture description in subsets (1) (–context) and (3) (+context) came from the same sample of four sentences per speaker. The sentences were distributed so that the same sentence from the same speaker would not appear in both –context and +context conditions in the same listener set. In the +context condition, the picture was visible to the listeners.

Intelligibility was rated as follows: Each listener was asked to orthographically transcribe the utterances from the picture description task and the STI-sentences. For the free narratives, listeners were asked to rate on a 6-point scale from 1 = the speech is not possible to understand, to 6 = the speech is very easy to understand.

### Fig. 1. Overview of set structure and tasks for listeners.

In each listener set, two randomly chosen sentences in each subset of transcription tasks and two extracts of narrative speech for scaled ratings were included twice for intra-judge agreement analysis. The repeated sentences from picture description and STI were placed at the end of their subsets and the repeated extracts of free narratives at the end of the listener set. Other than that, the same utterance from the same speaker did not appear twice in the same listener set in order to avoid familiarization effects.

The sound files were edited in Audacity 2.3.0. Before each sentence from picture description and STI, 2 s of silence was inserted, allowing the listeners to prepare. Peak volume was controlled and similar in all sound files. Microsoft PowerPoint was used as carrier software for the listener tasks. In total, the PowerPoint file for each listener set contained 64–65 slides with: (a) questions eliciting listener background data, (b) general and task-specific instructions, (c) two utterances from speakers not included in the test material for the purpose of training and so that listeners could adjust the sound level before commencing the actual listener task, and (d) one slide per sentence or free narrative extract.

### Collection of Listener Data

The listeners completed the tasks individually, using a laptop computer and high-quality headphones. Listeners typed background information, transcriptions, and ratings directly into the PowerPoint slides. The procedure was self-paced, and the listeners spent about 30–40 min completing the tasks. In addition to written instructions in the PowerPoint file, information was given orally.
The listeners were allowed to hear the free narrative extracts and utterances from picture description once, and the STI-sentences twice, since the latter were judged to put more demands on memory because of their length and the lack of context and predictability. In the instructions, listeners were encouraged to write what they perceived the person was saying, even if it was hard to identify and to mark unintelligible parts of an utterance with question marks or dots.

The 42 listeners were evenly distributed over the 6 listener sets. For the transcription tasks, there was a possible maximum of 28 listeners per speaker for each condition: −context, +context, and STI-sentences (4 sentences per speaker, distributed over 4 different listener sets, and 7 listeners per set). For 6 sentences in total (0.4%), transcription data was missing. The sentences with missing data were distributed over different conditions, listener sets, and speakers from both groups. Consequently, 6 speakers were, for one of the 3 conditions, judged by 27 instead of 28 listeners.

The extract of narrative speech from each speaker was randomly allocated to 2 different listener sets, resulting in a possible maximum of 14 scaled ratings of intelligibility per speaker. For scaled ratings, about 3.5% of the data were missing. This was distributed over 9 different speakers from both groups, who were rated by 13 instead of 14 listeners. One listener could not complete the scaled ratings because of technical problems with the larger sound files in this task. The rest of the missing data was probably due to listeners accidentally skipping a slide in the PowerPoint-file.

Reliability Analysis

For inter- and intra-judge agreement in the transcription of sentences and ratings of narrative speech, absolute agreement single measures intraclass correlations (ICCs) were calculated and interpreted according to Koo and Li [47]. Inter-judge agreement for the number of correctly transcribed words was in the moderate-to-good range for all four measures. Intra-judge agreement was excellent for the sentences from picture description with ICCs > 0.900, and moderate to good for the STI-sentences and scaled ratings. Both inter- and intra-judge agreement were somewhat lower for the scaled ratings than for the transcription scores. The ICCs from this analysis are presented in Table 2.

### Table 2. Inter- and intra-judge agreement in transcription of sentences and listener ratings

<table>
<thead>
<tr>
<th>Intelligibility measure</th>
<th>Inter-judge agreement&lt;sup&gt;1&lt;/sup&gt;</th>
<th>95% confidence interval</th>
<th>Intra-judge agreement&lt;sup&gt;2&lt;/sup&gt;</th>
<th>95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ICC</td>
<td>lower bound</td>
<td>upper bound</td>
<td>ICC</td>
</tr>
<tr>
<td>Sentence −context&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.729</td>
<td>0.650</td>
<td>0.804</td>
<td>0.940</td>
</tr>
<tr>
<td>Sentence +context&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.800</td>
<td>0.736</td>
<td>0.858</td>
<td>0.975</td>
</tr>
<tr>
<td>STI&lt;sup&gt;b&lt;/sup&gt;</td>
<td>0.735</td>
<td>0.656</td>
<td>0.808</td>
<td>0.774</td>
</tr>
<tr>
<td>Scaled ratings</td>
<td>0.596</td>
<td>0.443</td>
<td>0.752</td>
<td>0.713</td>
</tr>
</tbody>
</table>

<sup>1</sup> Intraclass correlations (ICCs) for inter-judge agreement were calculated based on a single measures, absolute agreement, one-way random-effects model. <sup>2</sup> Intraclass correlations (ICCs) for intra-judge agreement were calculated based on a single rater, absolute agreement, two-way mixed-effects model. <sup>a</sup> Sentences from picture description. <sup>b</sup> Sentences from the Swedish Test of Intelligibility (STI).
For measuring inter- and intra-judge agreement in target content words from the picture description task, percentage agreement was calculated. Inter-judge agreement was 84% in the −context condition and 90% in the +context condition. Intra-judge agreement was 95% in the −context condition and 97% in the +context condition.

Results

Results in the section below are presented at both group and individual level. An overview of intelligibility measures and results at group level is presented in Table 3, which also includes results from neurologically healthy speakers for comparison. Table 4 presents individual pooled scores from transcription tasks and scaled ratings of intelligibility, respectively, for the speakers with Parkinson’s disease. A table showing detailed individual results of the transcription scores from the different conditions is provided in the online supplementary material. In the following sections, the results connected to each of the goals for the study are presented together with graphs displaying the magnitude of individual differences between conditions.

Intelligibility in Self-Generated Utterances Compared to Utterances Elicited by Oral Reading

At group level, the intelligibility scores were, on average, 3.16% (SD = 15.25) lower for the self-generated picture description sentences −context than for the orally...
read STI-sentences, but 5.72% (SD = 18.93) higher for picture description sentences +context. The within-group differences between the conditions were not statistically significant. Results from pairwise testing of measures of picture description sentences versus STI-sentences are displayed in Table 5.

Intelligibility in Target Content Words Compared to the Whole Sentence

At group level, the intelligibility scores were, on average, 4% higher for the target content words compared to the whole sentences for both the −context and +context conditions, which was not statistically significant (Table 5). However, the data showed different patterns for speakers with higher intelligibility scores compared to speakers with lower scores (Fig. 3). For speakers 4–11, who all had a pooled intelligibility score higher than 70%, intelligibility in target content words was significantly higher than intelligibility in whole sentences (Z = −2.38;
Assessment of Intelligibility in Parkinson's Disease

For two speakers (speakers 2 and 3, both with pooled intelligibility scores below 70%), the pattern was reversed with noticeably higher intelligibility for whole sentences than for content words.

Effect of Listener Knowledge of Context on Intelligibility in Self-Generated Utterances

Intelligibility scores from picture description were, on average, 9% higher for the participants with Parkinson's disease when listeners knew the context, both for whole sentences ($Z = -2.84; p = 0.004$) and target content words ($Z = -2.19; p = 0.028$). These within-group differences were statistically significant. Results are shown in Table 5.

As to the individual results (Fig. 4), there was a positive effect of listener knowledge of context for most of the speakers, but of varying magnitude. When removing speaker 2, for whom the effect of context was particularly high, the within-group mean differences between –context and +context conditions decreased to 7.45% ($Z = -2.70; p = 0.007$) for sentences and 6.44% ($Z = -1.96; p = 0.050$) for target content words but were still statistically significant.

Correlations between Intelligibility Measures

Scores from picture description sentences –context correlated strongly ($r_s > 0.800$) with the other three measures, that is, picture description sentences +context, STI-sentences, and scaled ratings. The correlations between scaled ratings and picture description sentences +context and STI-sentences, respectively, were somewhat weaker, and for the speakers with Parkinson's disease in the moderate range ($r_s = 0.627$ and $r_s = 0.692$, respectively). Correlations between intelligibility measures are shown in Table 6.

Fig. 3. Differences between intelligibility scores for whole sentences and content words. Individual differences in intelligibility percentage scores between whole sentences from picture description and target content words in the sentences. The zero line represents the score for whole sentences. A positive value indicates that intelligibility is higher for content words than for the whole sentence. Speakers (Sp) are numbered based on the total pooled intelligibility percentage score, where Sp 1 has the lowest score (52%) and Sp 11 the highest (95%). –context = listener not provided with the context (picture), +context = listener provided with the context (picture).

Fig. 4. Effect of listener knowledge of context. Individual differences in intelligibility percentage scores from transcription of sentences and target content words from picture description. The zero line represents the score where listeners do not know the context (–context condition). A positive value indicates that intelligibility is higher when listeners know the context. Speakers (Sp) are numbered based on total pooled intelligibility percentage score, where Sp 1 has the lowest score (52%) and Sp 11 the highest (95%). Where a bar is missing (Sp 4 and Sp 9), there is no difference (value = 0) between the conditions.
Effects of Utterance Length

Group summaries of utterance length are presented in Table 7. There were no statistically significant group differences regarding the length of sentences or target content words.

At group level, the correlations between intelligibility and the length of utterance or target content words were weak ($r < 0.250$) and non-significant, except for a weak to moderate negative correlation with the number of syllables in sentences from STI for the participants with Parkinson’s disease ($r = -0.314; p = 0.038$), indicating lower intelligibility of sentences with a higher number of syllables. At individual level, the correlations between intelligibility scores and utterance length were in the weak to moderate range and non-significant for eight of the eleven speakers with Parkinson’s disease. For two of the speakers there were moderate to strong statistically significant negative correlations between intelligibility and both number of syllables and number of words in sentences: Speaker 1 (syllables: $r = -0.590; p = 0.043$; words: $r = -0.817; p = 0.001$) and speaker 2 (syllables: $r = -0.715; p = 0.009$; words: $r = -0.673; p = 0.016$). There was also a moderately strong negative correlation between the number of syllables and intelligibility for speaker 10 ($r = -0.582; p = 0.047$). Detailed information regarding individual results on correlations between intelligibility and utterance length are provided in the online supplementary material.

Discussion

In this study, we investigated the use of self-generated utterances based on picture description for assessing intelligibility in Swedish speakers with dysarthria consequent to Parkinson’s disease and in neurologically healthy persons. Miller et al. [4] discuss that scores from intelligibility tests can either over- or underestimate the actual
problem because of, on one hand, the lack of other cues for message interpretation and, on the other hand, the capacity for speakers with Parkinson’s disease to temporarily raise speech performance in a test situation. As is discussed in the literature on intelligibility [19, 21–23], different assessment methods have their strengths and weaknesses. There is probably not one single measure that in itself gives a true reflection of how easily a person can make him- or herself understood in different communicative settings and with different communication partners, but used in combination, different assessment methods can contribute information on different challenges and skills. The addition of self-generated utterances to the assessment of intelligibility could provide a more nuanced information on speech functionality, which could possibly lead to individually better-tailored interventions. Therefore, one aim of the present study was to try out transcription of self-generated utterances elicited through picture description as a method for assessment of intelligibility in speakers with Parkinson’s disease. We especially wanted to explore if adding transcription of self-generated utterances to the assessment has the potential to provide additional information about: (1) intelligibility in oral reading versus self-generated speech, (2) intelligibility of content words in self-generated sentences, and (3) effects of listener knowledge of context.

Scores from self-generated utterances in the −context condition were, on average, lower than for the orally read STI-sentences for both the speakers with Parkinson’s disease and the neurologically healthy speakers, but the difference was small (between 3% and 4%). For the picture description sentences +context and the nonsense STI-sentences, the difference was much smaller, with results showing positive as well as negative difference values for the different speakers. An explanation could be that contextual leads, on one hand, and performance effects in oral reading, on the other hand, may cancel each other out in terms of the resulting intelligibility measures, even though the impact of these two processes on the speech output would be qualitatively different, also with a possible inter-speaker variation. One speaker (speaker 2) showed a reversed pattern with extremely low STI-scores. These results could probably be explained by a combination of effects of utterance length and level of contextual leads. It should be noted that for this speaker, the very low score for the STI-sentences was not due to reading problems. The findings from the present study partly support findings from previous studies of performance effects in oral reading for speakers with Parkinson’s disease [12–15]. However, the occurrence of a case (speaker 2) with a score pattern deviating from the others also shows that there are individual exceptions.

The intelligibility for non-predictable content words in self-generated sentences tended to be higher than for the sentences in total, which ran counter to what was expected for the speakers with Parkinson’s disease based on previous research on intelligibility in different types of words [37] and common features in the speech in Parkinson’s disease, such as prosodic changes [2]. The results from the present study suggest that there might be a stronger tendency for listeners to focus on information-bearing words when speakers’ intelligibility is mildly-to-moderately affected. For speakers with more severely affected intelligibility, content words might not emerge as clearly. For speakers 2 and 3, the intelligibility scores for target content words were noticeably lower than for whole sentences. For these speakers, intelligibility in general was more affected, and one explanation could be that function words were comparatively easier for the listeners to detect due to their syntactic predictability [37]. For speakers where the transcription scores from picture description were markedly higher for target content words than for the whole sentences, the scaled listener ratings were also comparatively high. When content words in an utterance are easier to understand, listeners may judge the speech as generally more intelligible, even though they may not correctly identify all words.

Unlike Utianski et al. [38], we found that listeners knowing the context of an utterance significantly improved intelligibility for the speakers with Parkinson’s disease at group level, but the benefit of contextual leads in the form of a picture varied between individuals. This variation was not clearly related to overall levels of intelligibility and would possibly be best explained as an interaction between motor speech impairment, the individual’s way of linguistically expressing him- or herself, and the representativeness of the utterance in relation to the depicted context. For the neurologically healthy speakers, there was also a benefit of contextual leads, but the differences between the −context and +context conditions were smaller than for the speakers with Parkinson’s disease and statistically non-significant.

Overall, the neurologically healthy speakers showed a more homogenous pattern across tasks than the speakers with Parkinson’s disease, with generally high intelligibility scores, smaller standard deviations, and no statistically significant differences between conditions, which was an expected result also for the measures based on a transcription of self-generated utterances from picture description. However, the neurologically healthy speak-
ers’ intelligibility scores for the self-generated sentences —context were, on average, a few percentage points lower, and had a wider within-group variation than the scores for the orally read STI-sentences.

Another aim of the study was to investigate how intelligibility scores from self-generated utterances correlate with scores based on sentences from the STI, elicited by oral reading, and scaled ratings. In general, there were strong correlations between the scores from the different transcription tasks. According to Weismer [22], high correlations between different kinds of intelligibility measures would be expected, since they all are likely to be correlated with the overall severity of participants’ speech impairment. Correlations between scaled ratings and the transcription measures were somewhat lower. This may indicate that when making a more holistic judgement (as in a scaled rating), listeners perceive intelligibility somewhat differently than when focusing on segmental characteristics (as in transcription). The reliability analysis also showed lower agreement for scaled listener ratings than for the transcription tasks, especially for inter-judge agreement. This could be connected to a variation in listener criteria [25] and which aspects different listeners rate when they are asked to do a scaled rating of how understandable they perceive a sample of speech to be. The measure might not include only intelligibility of the speech signal, but maybe also elements of, for example, speech impairment severity, speech naturalness, and/or comprehensibility of the message. Woisard and Lepage [33] discuss that the variation in scaled ratings of intelligibility might be explained by differences in listener interpretation of the concept and the level for judgement: the syntax and morphology of the original message, the semantics, or the message integrated with the knowledge of the receiver.

Methodological Considerations

For this first-step study, we wanted to control and evaluate the effects of possible confounding variables. For example, sentences were extracted based on quite detailed criteria, and the number of listeners was fairly high. Therefore, the procedures used for the purposes of the present study are not immediately transferable to clinical practice. Precautions were taken in preparing the listener material to avoid effects of listener familiarization, such as distributing the speech material over different listener sets and the order of subsets of listener tasks. In this way, the same utterance from the same speaker did not appear twice in any listener set (except the repeated utterances used to explore intra-judge agreement). The same speaker could appear in more than one subset, but the risk of familiarization effects on the results is considered to be very low, since short samples of speech and passive familiarization, as in this study, do not seem to benefit intelligibility [5]. As previously mentioned, a challenge when judging intelligibility in self-generated utterances is knowing what a person with dysarthria has actually said because of the reduced control over content and utterance structure compared to oral reading tasks [24]. Since this study did not include any speakers with more severe dysarthria, we believe that there is a low risk that researcher misperceptions would bias the results. The 1st author, who performed the recordings of speakers, had long clinical experience as an SLP from working with persons with neurological communication disorders, including patients with Parkinson’s disease, and was familiar with the person’s speech before the picture description task from the initial interview and dysarthria assessment. Having the picture gave contextual support in the recording situation, where it was also possible to verify and ask for clarifications, if necessary. In addition, SLPs not linked with the study judged ambiguous utterances before collection and analysis of listener data.

To make transcription of self-generated sentences useful in clinic, some adaptations are suggested: In this study, we chose to use a rather complex picture with several ongoing events to obtain utterances from the participants’ description of the picture. Developing and selecting from a pool of single-action pictures could be more practical and would also aid reproducibility. The pool of pictures would need to be large to avoid familiarization effects. For the present study, the aim was to elicit utterances as close to the person’s natural speech as possible, but it would be possible to customize the content of pictures to assess different variables [27], such as effects of phonetic contrasts and/or word length [16]. To avoid the effects of semantic predictability, some unexpected elements could be included in each picture, for example, the use of an object in a non-typical way. As in the present study, the recording and the transcriptions, respectively, could be conducted by different persons. By doing that, it would be possible to ensure that the speaker’s utterances are correctly interpreted by verification already during the recording session, and it would also more easily enable assessment of the effect of listener knowledge of context. In the present study, the inter-judge agreement was quite strong, indicating that the measure could be reliable across listeners, which would also facilitate the transfer to clinical settings where a large number of listeners is rarely available.
A weakness of this pilot study is the low number of speakers, which limits generalization of the results. Another limitation is that the speakers in the present study had predominantly mild or moderate dysarthria. Future research on a larger sample, including speakers with more severe dysarthria, is therefore warranted. As regards the size of the speech sample from each speaker, the choices for the present study were based on the aims to have both a variety of speakers and tasks in the material and a reasonably large number of listeners per speaker, to enable inter- and intra-judge reliability analysis. Even though the number of utterances per speaker and task was limited, they still presented valid snapshots of each participants’ speech.

Control of effects of utterance length showed moderate-to-strong negative correlations between intelligibility scores and utterance length for three speakers, of which two had the most severely impaired intelligibility among the participants. Other than that, the correlations between utterance length and intelligibility in the present study were generally weak and did not clearly correspond to either the results from Beverly et al. [36] or Tjaden and Wilding [45]. The different types of speech materials and methodology in the studies also make comparison difficult.

Cognitive impairment is common in Parkinson’s disease [48]. In this study, a diagnosis of dementia was an exclusion criterion. Therefore, we could expect only minor, if any, cognitive impairment due to Parkinson’s disease in our sample, and there were no indications in the data that differences in MMSE-SR scores contributed to the noted individual variations in the results.

Clinical Implications

For optimal ecological validity, intelligibility assessment should, ideally, be based on real-life data from conversations in different situations and with different communication partners; however, such an approach would be very challenging in terms of clinical feasibility and reproducibility. Although picture description is not the same as real-life conversation, and despite reliability issues when transcribing self-generated utterances from speakers with dysarthria, picture description has the advantages of allowing for a more consistent and clinically feasible procedure than real-life conversation data does, and of being closer to natural speech than oral reading. Therefore, it can provide valuable complementary information on speech functionality to assessment methods based on sentences elicited by oral reading. Even though the correlations between the intelligibility scores from picture description and the STI-sentences were strong in the present study, the within-speaker variation confirms recommendations to use multiple methods when assessing intelligibility [49]. The results from the present pilot study indicate that this would be especially important for speakers with more severe speech difficulties and lower intelligibility, where intra-speaker differences between tasks tended to be of greater magnitude and show a more varied pattern. For speakers with more severe dysarthria or cognitive impairment than was the case in the present study, issues of reliability when quantifying intelligibility in self-generated utterances arise, but to add such a method to assessment methods based on pre-scripted sentences would still provide important qualitative information on abilities and strategies to get a listener to understand the content of a message.

Combining different assessment methods is more time consuming, but allows for the construction of multifaceted profiles, which aids in the identification of different intervention targets. One clear example of differences between intelligibility scores with dissimilar patterns in the present study are speakers 2 and 4. When looking more closely at the individual results from speaker 2, the effect of context appears to be especially strong. For this speaker, there was also a negative correlation between intelligibility and utterance length. The STI-sentences, which have no lexical contextual leads and were longer than the speaker’s self-generated utterances, had the lowest intelligibility score, while the results from the +context condition in the picture description task had the highest scores. For this speaker, the intelligibility scores for content words in the sentences were comparatively low. For speaker 4, there was a marked difference between intelligibility scores for target content words and whole sentences from picture descriptions. In the listeners’ scaled ratings, speaker 4 was also perceived as more intelligible than speaker 2. The clearer production of information-bearing words might have contributed to this difference, and this in turn could be due to individual speaking style or a compensatory strategy. These individual differences across tasks reveal components that increase intelligibility for the individual, suggesting different approaches for interventions to make use of these components and to detect speech behaviours that could need intervention for increased intelligibility. For example, possible intervention targets for speaker 2 could be strategies for topic announcement, as well as emphasizing key content-bearing parts of utterances. For speaker 4, a target could be to maintain a habit of emphasizing key content words (which this speaker already does) as Parkinson’s disease
progresses but to also balance this emphasis with attention to other utterance elements, such as word boundaries.

**Conclusion**

Multifaceted assessment can increase the ecological validity of speech evaluation since it captures a wider variety of demands made on speech and interactional abilities. Adding transcription of self-generated utterances, which more closely resemble everyday speech, to the assessment of intelligibility has the potential to provide additional information to assessment methods based on oral reading of pre-scripted sentences, and to inform the planning of interventions. The noted within-speaker variability also supports recommendations in the literature to use multiple methods and tasks when assessing intelligibility. Development and trial of intelligibility assessment material based on picture description for clinical use is suggested for future research. Future investigations should be based on a larger sample of speakers and include a higher proportion of speakers with severe dysarthria than was the case for this pilot study.

**Acknowledgments**

We would like to thank all our participants.

**References**


**Statement of Ethics**

Ethical approval for the study was obtained from the relevant regional Ethical Review Board (2015/446-31). All participants provided written informed consent prior to data collection.

**Conflict of Interest Statement**

The authors have no conflicts of interest to declare.

**Funding Sources**

This study was, in part, funded by the Foundation for Parkinson’s Research at Linköping University (LiU 2015-00194), the Regional Council of Östergötland (LiO 620581), and the Swedish Parkinson Foundation (1076/18).

**Author Contributions**

All authors contributed to the design and methodology of the work. Inga-Lena Johansson was responsible for collecting and analysing data, and for writing and editing the manuscript. Christina Samuelsson and Nicole Müller were responsible for supervising the project, as well as for reviewing and editing the manuscript. All authors have given their final approval of the version to be submitted and agree to be accountable for all aspects of the work.

**Data Availability Statement**

The data generated and analysed in this study are not publicly available, based on ethical approval and participant consent. Further enquiries can be directed to the corresponding author.
Assessment of Intelligibility in Parkinson’s Disease


