Chapter nine

Intelligibility of words in sentences

9.1 Introduction

In the preceding chapters we were concerned with the intelligibility of the smallest building blocks of language, i.e. the vowels and the consonants, in either meaningless sound sequences or in existing words and short phrases constructed such that the identification of target segments was not made (more) predictable by context. This is a listening situation that occurs only very rarely in everyday life. Normally sounds occur in meaningful words. Typically, when sounds occur in the context of a word in a sentence, the listener needs to get only a few of the constituent segments to piece the word together, using lexical redundancy. For instance, the last two sounds in the word elephant are perfectly predictable once the listener has heard /leɪf/. There are simply no other words in the English lexicon than elephant that begin with this sequence. When the target word is embedded in a meaningful context sentence, segments in short, monosyllabic words will also be predictable. If the listener misses the initial consonant in I heard the _at mew, the listener will know from his knowledge of the world that the entity that produced the mewing sound must be a cat rather than a rat (or bat or gnat), let alone a mat. In the present chapter we will deal with the intelligibility of meaningful words in several kinds of sentence contexts.

The first type of sentence context is a syntactically correct structure, but the words that are filled in the various slots in the structure do not yield a meaningful sequence. For instance, in The state sang by the long week, it is at least odd that an inanimate subject The state should perform an action normally only manageable by humans (i.e. singing); also, the choice of the preposition by would seem to be ungrammatical. These sentences were called Semantically Unpredictable Sentences (Benoît, Grice and Hazan, 1996; see also Chapter two) or just SUS sentences. They were originally constructed for the purpose of evaluating the quality of text-to-speech systems. The claim would be that the SUS test will discriminate in a highly sensitive way between small differences in speech quality, when the subjects are native listeners of the stimulus language. The test was not developed to discriminate excellent from not-so-excellent speakers and listeners.

The second type of test we used in our materials is the SPIN test, which stands for SPeech In Noise test (Kalikov, Stevens and Elliot, 1977). The SPIN test (see also Chapter two) requires listeners (patients with hearing loss, in the original application) to fill in the last word of a short sentence; the final word is either highly predictable (HP) from the preceding words in the sentence (e.g. She put her broken arm in a
sling) or not predictable from the context (low predictability, LP, e.g. We should consider the map). The SPIN LP sentences are more or less comparable with the SUS sentences in that the target words appear in grammatically correct word sequences, may benefit from the presence of a precursor utterance (phonetic adaptation to phonetic quality, melody, rhythmic structure and coarticulation) but not from any semantic constraints. Earlier comparisons of SUS sentences with SPIN sentences (Hazan and Shi, 1993), using normal English listeners, brought to light that the SUS sentences were much more difficult (12% correct on average) than the SPIN sentences (HP 84% correct, LP 48% correct).

We decided to include all three types of sentences in our test battery (i.e., SUS, SPIN-LP, SPIN-HP,) precisely because together they would seem to cover a very large range of listener abilities, large enough to adequately discriminate all nine combinations of speaker and listener nationalities in our study. More specifically, since the purpose of the SPIN audiology test was to discriminate between listeners from a wide range of hearing ability and that of the SUS test was to differentiate between better and poorer talking machines, one would expect therefore that the SPIN test will be rather more sensitive to differences between listeners, whilst the SUS test would be susceptible to differences between speakers.

I will now present the results of word recognition in sentences for each of the nine combinations of speaker and listener groups (Chinese, Dutch, and American). The results of the SUS sentences will be presented first (§ 9.2), followed by the SPIN sentences (§ 9.3). Within each test, I will first present the intelligibility scores in terms of percent correctly reported words. Here a word will be counted as incorrectly reported even if just one phoneme within the word was incorrectly reported. In a second analysis I will present a more refined scoring method where onsets, vocalic nuclei and codas are scored separately so that each target word may have a score of 0, 33, 67 or 100% correct. In between the two word recognition analyses, I will present and analyze the results for onsets, nuclei and codas separately. This latter breakdown of the data will afford a direct comparison with the vowel, consonant and cluster identification results in Chapters six, seven, and eight, respectively. The same sequence of results will then be repeated for the SPIN sentences.

9.2 Intelligibility in SUS sentences

Every listener heard 30 SUS sentences. These were evenly distributed over five different syntactic frames (see Chapter four, § 4.2.4) with each speaker (i.e., one male and one female Chinese, Dutch and American speaker) donating one sentence to each syntactic frame. Speakers were blocked over sentences such that any listener heard each sentence only once, and every speaker donated each sentence as often as any of the other speakers.

9.2.1 Overall result

A broad phonemic transcription was produced for all the stimulus (input) and response (output) forms. To this effect the orthographic input and output forms were
WORD INTELLIGIBILITY IN SENTENCES

converted to broad IPA by hand. In the case of the input forms this could be done efficiently, since the same words occurred in the same order for all of our 108 listeners; once the input list was transcribed it could simply be copied. The responses required much more work. All transcriptions were checked by an independent expert; whenever discrepancies were found between the transcribers, these were discussed and checked against a pronouncing dictionary (Kenyon and Knott, 1944). Stress marks were not included in the transcriptions of either input or output forms.

Figure 9.1 presents the overall percentages of correctly reproduced words broken down first by nationality of the listener and broken down further by nationality of the speaker.

![Image of Figure 9.1](image_url)

Figure 9.1. Percent correct word identification in SUS test for Chinese, Dutch and American listeners broken down by accent of speakers. Numbers above the bars indicate the subgroup membership as determined by the Scheffé procedure. Numerical values of means, N, SD and Se are included in Appendix A9.1.

The effect of listener nationality is highly significant by a two-way ANOVA with listener and speaker nationality as fixed factors, \( F(2, 312) = 669.0 \) (\( p < .001 \)). Post-hoc Scheffé tests reveal that the Chinese listeners (mean = 41% correct) performed more poorly than the Dutch (78%) and the American (79%) listeners, who did not differ from each other. There is a smaller effect of speaker nationality, \( F(2, 312) = 240.0 \) (\( p < .001 \)) by which Chinese speakers are poorest (52%), Dutch speakers are intermediate (70%) and Americans are best (77%). All three speaker nationalities

\(^1\) Unfortunately, the responses of one Chinese listener were missing, so that the number of valid listeners in this group was 35 instead of the nominal 36. This is reflected in the smaller number of degrees of freedom in the error terms in the ANOVAs.
differ from each other (Scheffé, p < .05). As was also observed in earlier chapters, the effect of listener nationality is appreciably stronger than that of speaker nationality (here roughly in a 3:1 ratio). As before, the speaker × listener interaction also reached significance, F (4, 312) = 45.9 (p < .001). The interaction is clearly the result of what we have called the interlanguage benefit in earlier chapters. For Dutch and American listeners, Chinese speakers are difficult to understand but Chinese listeners have word-recognition scores for fellow Chinese speakers which are not less than for the Dutch or American speakers. By the same token, Dutch listeners do relatively better for Dutch speakers than for speakers of other nationalities. Similarly, even American speakers have a small advantage when listening to their own speaker type.

Figure 9.2 lists the percentage of correct word recognition for each of the nominally 36 listeners in each nationality. The figure shows very clearly how sensitive the SUS test is. There is a clear gap in the distribution of the scores at 60%. Chinese listeners never obtain scores of 60% or more, while no Dutch or American ever gets a score below 60. There is virtually no difference between the Dutch and the American listener groups.

![Figure 9.2](image)

Figure 9.2. Correct identification (%) for words in SUS test by Chinese, Dutch and American listeners. Note that listener 17 is absent from the Chinese set of subjects (see also note 1).

### 9.2.2 Intelligibility of subsyllabic constituents

So far, we have merely analyzed the results in terms of the percentage of correctly recognized words. In order to obtain a more refined view of the specific difficulties we will now present percent correctly reported subsyllabic units, i.e., onsets, vocalic nuclei and codas. Since the consonant inventories of Chinese, Dutch and English differ less in size and complexity than the vowel inventories, we would predict that the effect of speaker and listener nationality will be greater for nuclei (vowels) than
for onsets (simplex initial consonants and clusters). Moreover, since both Dutch and English allow a wide variety of coda consonants and clusters, whilst Chinese only allows nasals in the coda, we predict large differences in percent correctly identified codas when speaker and/or listener nationality is Chinese. Once the scores are broken down by subsyllabic constituent, we may also derive a more refined overall word recognition score by counting correct onsets, nuclei and codas together.

Figure 9.3A-C presents the percentages of correctly identified onsets, nuclei and codas, respectively, broken down by nationality of listener and by nationality of speaker, as was done for overall word recognition in Figure 9.1. Figure 9.3D presents the composite word recognition scores, where each word can be recognized at 0, 33, 67 of 100% correct, depending on the number of subsyllabic constituents reported correctly.

Figure 9.3. Percent correctly identified onsets (A), vocalic nuclei (B), and codas (C) in word identification in SUS test for Chinese, Dutch and American listeners broken down by accent of speakers. Panel D plots a composite word-recognition score (see text).
Interestingly, the difference between the Chinese, Dutch and American listeners in Figure 9.3A is relatively minor (80, 95, 95% correct), \( F(2, 312) = 302.2 \) (\( p < .001 \)), for the onsets. It is greater for vocalic nuclei (Figure 9.3B, 68, 83, 86% correct), \( F(2, 312) = 469.1 \) (\( p < .001 \)), and greatest for the codas (Figure 9.3C, 56, 84, 85% correct), \( F(2, 312) = 527.1 \) (\( p < .001 \)). The prediction formulated above, i.e. that the difference between the three listener nationalities would be smallest in the onsets, intermediate in nuclei and greatest in the coda, is borne out by these data.

Observe that there is a striking resemblance between Figure 9.3A and Figures 7.1 (simplex onset consonants) and 8.1 (complex onsets). In all three figures the Chinese listeners exhibit considerable interlanguage benefit, which graphically shows up in the poorer identification rates for Dutch-accented onsets. This would show that the detailed tests using nonsense sound sequences in Chapters seven and eight make valid predictions of the listeners’ behavior in meaningful words. A more detailed error analysis may be done to examine to what extent the consonants and clusters that were problematic in the nonsense materials are also problematic in the meaningful words.

Figure 9.3D shows that the more refined scoring mechanism is beneficial to the poorer speaker and listener groups. The percentages of correct scores for the Chinese listeners are elevated from ca. 40% to ca. 60%. Similarly, the scores for the Chinese speakers are raised from 60 to 80% when the listeners are either Dutch or American. Clearly, the composite word recognition scores discriminate less effectively between poorer and better combinations of speaker and listener groups.

9.3 Intelligibility in SPIN sentences

9.3.1 About the SPIN test

The SPIN test (Speech In Noise) was developed as a diagnostic tool to determine the severity of hearing loss in audiological settings. The article in which the concept of the SPIN test was introduced (Kalikov et al., 1977) mentions that the test had not been administered systematically to patients but data were presented to a normal-hearing reference group of American listeners. These data can be used as a background against which some of our own data can be gauged. SPIN sentences should be administered at various signal-to-noise levels. In our application we did not do this, as we noted in pilot versions of our test that the range of intelligibility across the various speaker and listener types was more or less fully covered; had we presented stimuli in noise, some of our listener groups would not have understood a single word.

The SPIN test presents sentence-final target words in high-predictability (HP) and in low-predictability (LP) contexts (see introduction). In the LP contexts the results should be roughly similar to those obtained in the SUS sentences. In both type of tests, the target words have to be understood purely from bottom-up acoustic information contained in the word itself; syntactic and semantic cues in the preceding context are useless. In the HP sentences, the words in the preceding context strongly constrain the identity of the sentence-final target word. In this condition, the SPIN test comes rather close to real-life speech recognition, where the
outcome of the processing task is the result of interaction between acoustic bottom-up information and top-down semantic and syntactic information. It seems a reasonable hypothesis that the interaction between the two information sources makes heavier demands on the listener, so that the native listeners will benefit substantially from the contextual information but that the non-native listeners will be hindered by the dual-processing task – having to attend to two non-automatized processing tasks at the same time and not doing a good job on either.

9.3.2 Overall word recognition in SPIN sentences

We will first present the results in terms of overall word recognition, once across both predictability conditions, and then separately for LP and HP sentences. In this part of the data presentation a word will be counted as an error if any component of it was not correctly reported by the listener, whether a coda consonant, a vocalic nucleus of some part of the coda.

As before, a broad phonemic transcription was produced for all the stimulus (input) and response (output) forms. All the target words were monosyllabic. However, in just a few cases listeners reported a two-syllabic word, e.g. bet was twice reported as better and lane as today. In such cases the segments of input and output forms were aligned manually such that the best match was obtained for an onset, nucleus and a coda. In the examples just given the first three segments of better were aligned with bet (also respecting the stress location); in today the /d/ was aligned with /l/ of lane (error), the two stressed vowels with each other (correct) and an empty coda was matched with the /n/. Non-aligned (extra) phonemes were not included in the analysis. Differences between the aligned input and output transcription were detected automatically; the scoring of the responses was done by computer. When even a single mismatch was found between input and output form, the entire word was scored as an error. In other words, every single segment in the word had to be reported correctly or else the word was not counted as a correct response. This is a very strict scoring principle. A more sophisticated scoring system in which errors in onsets, nuclei and codas were counted separately will be presented in a later section (§ 9.3.3).

Figure 9.4 presents the percentages of correctly recognized target words as defined here, broken down by nationality of listener and of speaker. The data have been accumulated over the two predictability conditions.

The data in Figure 9.4 were subjected to a three-way ANOVA with predictability (LP versus HP) of the targets, nationality of speaker and nationality of listener as fixed effects. The effect of listener was largest, $F(2, 630) = 807.6$ ($p < .001$), with Chinese listeners scoring 27% correct word recognition, Dutch listeners 63% and Americans 77%. All three listener groups differed significantly from each other (Scheffé, $p < .05$). A smaller effect was obtained for speaker nationality, with Chinese speakers performing significantly poorer (32%) than the Dutch and American speakers (both at 67%), $F(2, 630) = 500.4$ ($p < .001$). The effect of contextual predictability is much smaller, with 52 versus 60% correct words for LP and HP, $F(1, 630) = 58.8$ ($p < .001$). There was significant interaction between speaker and listener nationality, $F(4, 630) = 71.7$ ($p < .001$), which to some
extent reflects interlanguage or native language benefit. However, there is one remarkable instance of foreign-language benefit: the Chinese listeners perform significantly better when the speakers are Dutch than when the speakers are either Chinese or American. Possibly, the Dutch non-natives speak more slowly and deliberately than the American native speakers, which may have helped the Chinese listeners to get more useful information from the signal than with other speaker nationalities.

Figure 9.4. Percent correct word identification in SPIN test for Chinese, Dutch and American listeners broken down by accent of speakers. Numbers above the bars indicate the subgroup membership as determined by the Scheffé procedure. Numerical values of means, N, SD and Se are included in Appendix A9.2.

There is also significant interaction between the predictability condition of the targets and listener nationality (but not with speaker nationality), \( F(2, 630) = 22.6 \) (\( p < .001 \)). We will analyze the interaction in the next paragraph. Also the three-way interaction was significant, \( F(4, 630) = 18.0 \) (\( p < .001 \)). We will first analyze the two-way interaction (in Figure 9.3), and then we will analyze the three-way interaction by presenting the results for LP and HP separately (in figure 9.6A-B).

Figure 9.5 shows the interaction between predictability and listener nationality in detail. The figure shows that there is no effect of contextual predictability for the non-native listening groups, whether Chinese or Dutch. However, the difference is significant for the American listeners; here HP targets get better recognition scores than their LP counterparts. It seems, therefore, as if only the Americans profit from the contextual information. This would be in line with our suggestion above that non-native listeners do not recognize enough of the context to use it to their advantage.
We will now present the word recognition scores for the LP and HP conditions separately. This is done in figure 9.6A-B.²

Comparing the two panels by listener nationality, we may observe, first of all, that the Chinese listeners benefit from HP words somewhat but only if the speakers are American. Also the gain in percent correct is counteracted by a small loss of intelligibility in the HP utterances of Chinese and Dutch speakers. The Dutch listeners have no advantage of HP words at all. Apparently, they fail to use the semantic information contained in the meaningful context preceding the targets. The American listeners present an altogether different configuration of scores. If the speakers are American it does not really matter whether the words are LP (95% correct) or HP (97% correct). The quality of the pronunciation is such that recognition is close to ceiling in both conditions; there is no room for improvement due to HP. However, when the speakers are non-native, the pronunciation is relatively poor, in fact very much poorer for Chinese speakers (37% correct) and rather poorer for Dutch speakers (77%). When these speakers are tested with HP words, the Americans get so much useful information from the context that they improve their recognition scores by roughly 25 percent for Dutch speakers and by 20 percent for Chinese speakers. So, the significant three-way interaction mentioned

² There are minor differences in the word recognition scores in figures 9.6A-B and earlier reports of the data (e.g. Wang and Van Heuven, 2005). The reason for the small discrepancies is that some errors in the database (wrong alignments of input and output transcriptions were corrected in the present final analysis.
above is due to the fact that contextual information is only used by native listeners, and only if there is room for improvement, that is, when the speakers are foreign.

Figure 9.6. Percentage of correct word identification in SPIN test for Chinese, Dutch and American listeners broken down by accent of speakers, for low-predictability words (panel A) and for high-predictability words (panel B). Numbers above the bars indicate the subgroup membership as determined by the Scheffé procedure. Numerical values of means, N, SD and Se are included in Appendix A9.2.

9.3.3 Recognition of subsyllabic units in SPIN sentences

We will now examine the intelligibility of the subsyllabic components of the LP and HP target words. Figures 9.7A-B-C present percent correctly reported onsets, nuclei and codas for the LP words; figure 9.8A-B-C will do the same for the HP words.

As also appeared in Figure 9.3A for the SUS sentences, the difference between the Chinese, Dutch and American listeners in Figure 9.6A is relatively minor (73, 92, 95% correct), $F(2, 630) = 263.1$ (p < .001; all listener groups differ significantly, Scheffé) for the onsets. The difference is greater for vocalic nuclei (Figure 9.6B, 56, 78, 83% correct), $F(2, 630) = 286.4$ (p < .001; all listener groups differ significantly, Scheffé), and greatest for the codas (Figure 9.3C, 47, 71, 81% correct), $F(2, 630) = 354.9$ (p < .001; all listener groups differ significantly, Scheffé). Again, the hypothesis that the difference between the three listener nationalities would be smallest in the onsets, intermediate in nuclei and greatest in the coda (see also § 9.2.3) is borne out by these data. The resemblance also shows that the SUS sentences are highly comparable to the SPIN-LP sentences.
Figure 9.7. Percent correctly identified onsets (A), vocalic nuclei (B), and codas (C) in word identification in SPIN-LP test for Chinese, Dutch and American listeners broken down by accent of speakers. Panel D plots a composite word-recognition score (see text).
9.4 Conclusions

In this chapter we focused on the intelligibility of words spoken in the context of sentences rather than on the intelligibility of individual vowels and consonants in informationless contexts. Two types of sentence test were used: SUS and SPIN. The first test presented words in syntactically correct but semantically anomalous sentences, in which the function words correctly constrained the content words in terms of part of speech category but not in terms of meaning. One would expect words in such sentences to be difficult to understand. The second test contained syntactically and semantically correct sentences, which were constructed such that the sentence-final target word was either highly predictable from the preceding context (HP) or not. In the low-predictability sentences (LP) the context was neutral as to the identity of the targets, i.e. they were neither made more nor less predictable than when they had been presented as citation forms. All else being equal, the order of difficulty between the three types of sentences would be SUS > SPIN-LP > SPIN-
Table 9.1. Word recognition scores for SUS, SPIN-LP and SPIN-HP sentences broken down by nationality of listener and of speaker. See appendices A9.1-2 for number of listeners, and values of SD and Se.

<table>
<thead>
<tr>
<th>Nationality of</th>
<th></th>
<th>SUS scores by</th>
<th></th>
<th>SPIN scores</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Listener</td>
<td>Speaker</td>
<td>word</td>
<td>sentence</td>
<td>LP</td>
<td>HP</td>
</tr>
<tr>
<td>Chinese</td>
<td>Chinese</td>
<td>39</td>
<td>5</td>
<td>19</td>
<td>17</td>
</tr>
<tr>
<td>Chinese</td>
<td>Dutch</td>
<td>39</td>
<td>6</td>
<td>39</td>
<td>38</td>
</tr>
<tr>
<td>Chinese</td>
<td>American</td>
<td>44</td>
<td>5</td>
<td>18</td>
<td>32</td>
</tr>
<tr>
<td>Dutch</td>
<td>Chinese</td>
<td>57</td>
<td>17</td>
<td>27</td>
<td>33</td>
</tr>
<tr>
<td>Dutch</td>
<td>Dutch</td>
<td>86</td>
<td>60</td>
<td>81</td>
<td>76</td>
</tr>
<tr>
<td>Dutch</td>
<td>American</td>
<td>91</td>
<td>71</td>
<td>78</td>
<td>85</td>
</tr>
<tr>
<td>American</td>
<td>Chinese</td>
<td>60</td>
<td>18</td>
<td>39</td>
<td>58</td>
</tr>
<tr>
<td>American</td>
<td>Dutch</td>
<td>83</td>
<td>52</td>
<td>68</td>
<td>99</td>
</tr>
<tr>
<td>American</td>
<td>American</td>
<td>96</td>
<td>85</td>
<td>95</td>
<td>99</td>
</tr>
<tr>
<td><strong>Overall</strong></td>
<td>66</td>
<td>36</td>
<td>52</td>
<td>60</td>
<td></td>
</tr>
</tbody>
</table>

The table shows that the overall prediction does not hold: the SUS sentences are the easiest type. However, within the two types of SPIN sentences the prediction is correct: words in HP sentences are easier than words in LP sentences but the difference is rather small (but significant, cf. § 9.3.2). Reasons why the SUS sentences obtained better scores than either of the SPIN sentence types will be discussed in § 9.5.

The overall word recognition scores tend to be more extreme for the SPIN sentences than for the SUS sentences. The least and most favorable speaker/listener combinations in the SUS test are Chinese/Chinese and American/American with 39 and 96% correct, respectively. The comparable numbers for the SPIN-LP test are 19 and 95%, and for the SPIN-HP test 17 and 99. The discriminatory power of the various types of tests used in this dissertation will be examined in more detail in the next chapter. For now it will suffice to say that tests seem to discriminate better as they come closer to real-life speech perception, i.e. words in normally constrained, meaningful sentences. Interestingly, although the SPIN sentences were developed as audiological test materials to be presented in a range of signal-to-noise ratios, no degradation by added noise was needed in order to create a sufficiently wide range of scores in the present application of the test. Clearly, the suboptimal performance of the non-native speakers and listeners compensated for the absence of added noise.

For all three types of test (SUS, SPIN-LP, SPIN-HP) we find that the largest effect is that of listener nationality. It is stronger than the effect of speaker nationality by a factor 3. For both listener and speaker effects we find that the Americans obtain the highest scores, closely followed by the Dutch nationals, while the Chinese subjects performed much more poorly. The effects of context, as determined by comparing the SPIN-LP and HP sentences, are generally minimal,
except for American native listeners; only native listeners use the information contained in earlier words in the sentence to predict the identity of the sentence-final target word.

Again we observed clear effects of the interlanguage benefit, showing that listeners who hear speakers of their own nationality obtain better scores than when they are exposed to speech of speakers from a different nationality.

In Chapter three we predicted that coda consonants would present problems especially for Chinese speakers and listeners, as Chinese does not have any coda consonants, except the nasals /m/ and /n/. In the earlier chapters on the production and perception of vowels, consonants and clusters, no materials were included on codas. The only possibility to test the effects of onset versus coda consonants and clusters is to examine the scores in the present word recognition tests. The results in Figures 9.3 for the SUS sentences and 9.6 and 9.7 show, for SPIN-LP and SPIN-HP sentences, respectively, that the greatest differentiation between the Chinese and the other listener nationalities is found in the coda consonants; differentiation is somewhat poorer in the vowels and least in the onsets.

The last conclusion we will draw from this chapter is that not much is gained by computing a partial word recognition score based on correct identification of sub-word constituents. Generally, the scores for partial word recognition show the same tendencies as those for the constituent parts; moreover, when compared with the overall word recognition scores the results show the same order among the nine speaker/listener combinations but in a more compressed range, i.e. with poorer differentiation among the nine combinations.

9.5 Discussion

There is a remarkable discrepancy between our results and those reported by Hazan and Shi (1993) (see also § 9.1). In both studies a comparison can be made of the results obtained with SUS sentences and with SPIN sentences. Hazan and Shi found word recognition scores of 12, 48 and 84 percent correct for SUS, SPIN-LP and SPIN-HP sentences, respectively. My results reveal not the slightest difference between the scores on the SUS sentences and those on the SPIN-LP materials. Moreover, although the overall effect of LP versus HP sentences in the SPIN test is preserved in my study, the effect of context was only found for American listeners when the speakers were non-native.

Hazan and Shi (1993) recorded the materials from one male British English speaker and presented the materials to 50 native listeners. The materials were presented with a signal to noise ratio of 6 dB. It is possible, therefore, that the degradation due to the poorer signal-to-noise ratio (SNR) caused the enormous differentiation between the three tests in Hazan and Shi. We presented all our materials in quiet. As a result percent correct word recognition is close to ceiling in all three tests – but only if American native listeners respond to American speakers. When our speakers and/or listeners are non-native, the scores are rather more in the middle of the range. However, in our edition of the tests, there was virtually no difference between the LP and the HP word in the SPIN sentences (except when American listeners responded to American speakers) and the SUS sentences were
WORD INTELLIGENCE IN SENTENCES

some 10% better than the SPIN sentences for all conditions involving a non-native party. We must assume that the relative ease of the SUS test was caused by the way we presented the materials, i.e. not just once but repeatedly using a gating method incrementing the utterance in word-sized chunks.

The most important reason, however, why the mean SUS scores in Hazan and Shi were so low would seem to lie in the fact that these authors used the sentence as the scoring unit, whereas I computed word-recognition scores. In Hazan and Shi (1993), if even one word in a SUS sentence was wrong, then the entire sentence was wrong. In order to check whether my results would be more comparable to those of Hazan and Shi, I recomputed the SUS scores using the sentence as the scoring unit. The results in terms of the sentence-based scores have been listed in Table 9.1, along with the word recognition scores, as well as in Appendix A9.1.

Overall, the SUS scores drop from 66 to 36% when the sentence is used as the scoring unit instead of the word. As a result of this, the SUS scores are closer to those reported by Hazan and Shi (18% correct sentence recognition) but they are still considerably better. Moreover, the discriminatory power of the SUS sentence-based scores is better than that of the word-based scores. This property of the SUS test has been reported earlier by the designers of the SUS test (Benoît et al., 1996: 388). I will come back to the issue of discriminatory power of the tests used in my research in Chapter ten.