# MULTILINGUAL SPEECH RECOGNITION AT DRAGON SYSTEMS

*J. Barnett, A. Corrada, G. Gao, L. Gillick, Y. Ito, S. Lowe, L. Manganaro, and B. Peskin*

Dragon Systems, Newton, MA, USA

## ABSTRACT

This paper reports on some Dragon Systems' experiments with multilingual large vocabulary speech recognition, both for its discrete-word product DragonDictate® for Windows Version 1.0 and for its speaker-independent continuous speech research systems. The experiments in discrete word recognition involve English, French, German, Italian, and Spanish. The tests show significant, but not overwhelming, differences between the languages, with French being the hardest language to recognize and Italian being the easiest. The continuous speech experiments involve the Ricardo and CallHome corpora of conversational telephone speech, and show such high word error rates that no language-specific differences emerge. However, experiments with English Switchboard and CallHome recognition indicate that improved recognition technology and larger amounts of training data can improve accuracy substantially. We therefore expect that future multilingual LVCSR experiments will be more illuminating.

## 1. INTRODUCTION

Dragon Systems' large-vocabulary discrete-word dictation system, DragonDictate® for Windows (DDWin), is available in six languages: English (both US and British), French, German, Italian, Spanish, and Swedish. In addition, Dragon has large-vocabulary continuous-speech research systems in four languages: English, Spanish, Japanese, and Mandarin. The core speech recognition technology in both the product and research systems is language-independent, so that language-specific information is restricted to the tables of acoustic- and language-modeling data. Since the underlying recognition technology is being held constant, variations in recognition performance among languages will reflect differences in the languages themselves, as long as the system is trained on similar quantities of data of comparable quality in the various languages. In this paper we attempt to assess the difficulty of recognition in various languages by reporting on the comparative performance across languages of both the discrete-word dictation product line and the continuous speech research systems.

## 2. DISCRETE WORD RECOGNITION IN FIVE EUROPEAN LANGUAGES

In this section, we discuss the performance of the DDWin 1.0 large-vocabulary dictation system in English, French, German, Italian, and Spanish. To hold the recognition task roughly constant across the languages, we chose five texts which had already been translated. The texts represented a variety of styles, from Hegel's "Phenomenology of Mind" to Dragon's user's manual, and ranged in length from 1250 to 3700 words. Each text was dictated in each of the languages by four native speakers, two males and two females, who, with one minor exception, had not been part of the training data. The resulting scripts were run through the language-specific versions of DDWin using a 60,000-word active vocabulary. Since DDWin is speaker-adaptive, we started with a speaker-independent version of the system in each case, and permuted the order of the scripts, so that each text was tested using the other four as adaptation data. The data presented below therefore represent well, but not optimally, adapted performance. In Table 1, the first row represents average word accuracy over the five scripts, while the second row gives the individual speakers' range. (For more detail, including the amounts of training data, see [1].)

| Language | Word Accuracy | Individuals' Range |
|---|---|---|
| English | 90.68% | 89%-92% |
| French | 86.54% | 85%-87% |
| German | 86.74% | 85%-88% |
| Italian | 91.74% | 90%-93% |
| Spanish | 89.45% | 86%-91% |

**Table 1**: Mean Word Accuracy and Individuals' Range

An F-test shows that the differences in word accuracy are statistically significant at .001, indicating that isolated word recognition is more difficult in some of the languages than in others. One of the most striking differences among the languages is the variation in out-of-vocabulary (OOV) rates. As Table 2 shows, English has the lowest OOV rate, while German has the highest.

| Language | OOV Rate |
|---|---|
| English | .53% |
| French | 2.06% |
| German | 4.28% |
| Italian | 2.26% |
| Spanish | 2.35% |

**Table 2**: Out of Vocabulary rates

The differences in OOV rates are clearly due to the fact that the other languages are more highly inflected than English and that the recognizer treats each inflected form as a separate word. The German number is deceptively high, however, since it includes

many noun compounds that could have been dictated as individual words (and then combined using a formatting command). If we remove the unknown words that were either capitalized forms of words that were in the 60K lexicon in lower case (i.e. nominalized adjectives) and nouns that were compounds of words that were in the 60K lexicon, the German OOV rate comes down to 2.41%, which is in line with the rates for the other three inflected languages, but still much higher than the rate for English. Clearly, vocabulary coverage is an issue for inflected languages.

To get a handle on the acoustic differences between languages, we removed the OOV words from consideration and computed word accuracy for the words in the 60K active lexicon. Table 3 shows the results, which are still statistically significant, but only at .01.

| Language | Word Accuracy |
|---|---|
| English | 91.15% |
| French | 88.33% |
| German | 90.60% |
| Italian | 93.81% |
| Spanish | 91.58% |

**Table 3**: In-Vocabulary Accuracy

On the assumption that the language models in the various languages are similar, the performance differences shown in Table 3 should reflect acoustic properties of the languages. In particular, French appears to be the hardest language to recognize, while Italian is the easiest. To get a better idea of the acoustic difficulties, we looked at the 500 most common errors in each language. In discrete speech, all errors are substitutions, so we calculated the number of substitutions that involved homophones. Table 4 shows the percentage of errors involving homophone pairs among those errors caused by the top 500 error pairs (in effect, the errors are weighted by frequency).

| Language | Homophone Errors |
|---|---|
| English | 24.5% |
| French | 73.4% |
| German | 21.9% |
| Italian | 16.7% |
| Spanish | 24.6% |

**Table 4**: Homophone Errors among Top 500 Errors

The French number is astonishingly high when compared to those of the other four languages. Italian, on the other hand, has a relatively low rate of homophony. Moving a bit further, we considered 'near homophones', which we defined to be words that differed only in their final consonant (including cases where one word lacked the consonant altogether). The acoustic confusability of German shows up here, since 16% of the errors involved near homophones (for example, many common German inflectional variants differ only in final 'm' vs. 'n'). In contrast, in Spanish the percentage of near homophones was 9%, while in Italian it was down to 2%. These isolated-word experiments leave out all cross-word effects, but they seem to indicate that French is a relatively hard language to recognize, while Italian has a low degree of acoustic confusability and is thus relatively easy to recognize.

## 3. Continuous Speech Recognition

Our experiments with multilingual large-vocabulary continuous speech recognition have involved conversational telephone speech, which is a very different and much more difficult task than the isolated word dictation described above. Although the continuous recognition engine is completely language-independent, the high error rates on this task make it hard to detect differences between the languages. One set of experiments involved the Ricardo corpus of elicited telephone monologues which Dragon Systems collected in English, Spanish, Japanese, and Mandarin. (All the corpora mentioned in this section are available through the Linguistic Data Consortium.) Data was collected from approximately 40 speakers in each language. The subjects were prompted with a set of 21 questions and given up to a minute to respond to each one. The result was an average of about 15 minutes of speech per person. Roughly 7-8 hours were used for training, while two hours were used for testing. Table 5 shows the average word accuracy for the four languages. (N.B. the results for Mandarin are word-accuracy, not character-accuracy, as is sometimes reported.)

| Language | Word Accuracy |
|---|---|
| English | 40% |
| Spanish | 40% |
| Japanese | 34% |
| Mandarin | 38% |

**Table 5**: Ricardo word accuracy

Another set of experiments involved the CallHome corpus of two-sided telephone conversations in the same four languages. This corpus consists of spontaneous conversations, usually between friends or family members, recorded over long-distance phone lines. 100 two-sided conversations were used in each language, with 80 conversations (13 hours) used for training data, while 20 (about 3 hours) were held out as development test data. Table 6 shows the average word accuracy across the languages.

| Language | Word Accuracy |
|---|---|
| English | NA |
| Spanish | 24% |
| Japanese | 22% |
| Mandarin | 25% |

**Table 6**: CallHome word accuracy

An F-test shows that the differences in the CallHome results are not statistically significant. Note, however, that the different test speakers are engaged in different conversations, so that some of the speaker-to-speaker variability may come from differences in topic. Therefore, unlike the isolated word test, the identity of the speaker is not the only source of variation. In any case, the differences among the languages are much lower than one would expect given that Japanese and Mandarin are much more different from English and Spanish and from each other than the five European languages discussed above. We assume that the problem is that our models are not yet sharp enough to capture language-specific subtleties. For example, the Mandarin recognizer has no treatment of tone, yet the Mandarin CallHome recognition is slightly better than that in Spanish and Japanese, which lack tones.

Differences in vocabulary size and coverage do show up clearly across the languages. This is not surprising because the morphology of the language is substantially the same in written and spoken styles (though one might expect that the number of different inflected forms would be lower in conversation than in more formal styles). Spanish and Japanese, which have rich morphology, have lower coverage with larger lexicons than English and Mandarin. Table 7 shows the number of words in the lexicon and the out-of-vocabulary rate for Ricardo. Table 8 gives lexicon size, OOV rate and perplexity (using a simple bigram language model) for the CallHome development test data. As a further indication of the coarseness of the recognition, note that the lower perplexity of the Japanese test data did not translate into higher word accuracy.

| Language | Lexicon Size | OOV Rate |
|---|---|---|
| English | 5700 | 5.5% |
| Spanish | 6700 | 11.7% |
| Japanese | 7200 | 8.3% |
| Mandarin | 5900 | 7.0% |

**Table 7**: Ricardo lexicon size (number of words, excluding multiple pronunciations) and OOV rates

| Language | Lexicon Size | OOV Rate | Perplexity |
|---|---|---|---|
| Spanish | 9285 | 5.6% | 155 |
| Japanese | 9443 | 6.5% | 115 |
| Mandarin | 6213 | 3.5% | 149 |

**Table 8**: CallHome lexicon size (number of words, not counting multiple pronunciations), development test OOV rate and perplexity (bigram model).

The question now arises of why the Ricardo and CallHome tasks are so hard. Continuous speech recognition is not a problem in and of itself. For example, speaker-independent recognition accuracy on the Wall Street Journal task, which involves carefully read speech recorded with a high quality microphone, is above 90% at Dragon Systems and at other sites. On this test, at least, continuous speech recognition achieves levels of performance that are as good or better than the isolated-word results mentioned above. (Note, however, that the Wall Street Journal domain is narrower than the range of subjects covered by the texts used in the isolated word experiments.)

Various theories have been presented for why conversational speech is difficult. [2] presents evidence that the relatively high frequency of short words is an important factor in the degradation of recognition performance in conversational speech. The abundance of short words poses a problem both because they are hard to recognize and because the increased frequency of word endings leads to a higher effective branching factor. In addition [3] indicates that spontaneous speech is much harder to recognize than a read version of the same text. We would add another suspect to the list, namely lack of training data. For example, Switchboard is a difficult corpus of (English) conversational telephone speech, but Dragon's word accuracy has gone from less than 25% in 1993 to more than 65% in 1996. While this is in large part a reflection of improved technology, it is also true that we now train on over ten times as much data as in 1993. The 170 hours of Switchboard data that we now use is a marked contrast to the 13 hours available for each of the CallHome languages.

Table 9 shows the interaction of technological advances and increased training data in improving Switchboard recognition. The first column, labeled 'Baseline System', shows results for a simple system (gender-independent models, bigram language model) that is analogous to the one used in the CallHome evaluations reported in Table 6. The rightmost column, labeled 'Improved System', reports results for a system that used such improved features as rapid adaptation and speaker normalization. The columns show the effect of training these systems on increasing amounts of Switchboard data. (The language is English in all cases.) Reading down the columns shows that increasing the amount of training data from 13 hours to 60 hours adds close to 5% to word accuracy for the baseline, and even in the advanced system, going from 60 to 170 hours adds another 2%. But holding the amount of data constant at 60 hours and improving the technology yields even bigger improvements, as the middle row shows.

The important thing to note about Table 9 is that the CallHome results in Table 6 represent the amount of data and the level of technology shown in the upper left corner of Table 9, so that we can expect substantial improvements in CallHome results in all languages by increasing the amount of data and improving the recognition technology. Note, however, that the 43% accuracy on Switchboard shown in the upper left corner of Table 9 is still substantially better than the CallHome performance in all three languages. This may be due to the fact that the test was artificially easy, among other reasons because it involved only 10 Switchboard topics, instead of the full set of 70. However, this

relative improvement may also indicate that Switchboard is an easier corpus than CallHome.

| Training Data | Technology | |
|---|---|---|
| | Baseline System | Improved System |
| 13 hours | 43.2% | NA |
| 60 hours | 48.0% | 61.2% |
| 170 hours | NA | 63.2% |

**Table 9**: Effects of improved technology and increased training data (English)

Our current best system, which uses a more sophisticated language model than either of those in Table 9, gets 65% accuracy on this same task when trained with 170 hours of data. Using this same system, which was trained only on Switchboard data, we ran another experiment comparing the recognition of Switchboard and CallHome English. The results are shown in Table 10. This test was harder than the one reported in Table 9 for various reasons (one of the most important being that it required segmenting the speech stream to extract utterances), so the Switchboard performance is a bit lower than 65%, but the technology in question is slightly better than the system in the lower right-hand corner of Table 9.

| Corpus | Word Accuracy |
|---|---|
| Switchboard | 61% |
| CallHome | 50% |

**Table 10**: Comparison of Switchboard and CallHome Accuracy (English)

As before, we find that performance on Switchboard is better than on CallHome, again in part because Switchboard may be an easier corpus, but also because this recognizer was trained only on Switchboard data. Nonetheless, the 50% word accuracy on English CallHome is double the performance on any of the three languages shown in Table 6. It is clear that this improvement is not due to English being somehow easier to recognize (cf. Tables 1, 3), but rather due to the combined effect of increased training data and better recognition technology. Under similar test circumstances, we would expect similar results on Spanish, Japanese, and Mandarin. Therefore, if the amount of training data is increased, we believe that improvements in technology will lift the next generation of CallHome tests to a performance level at which cross-language comparisons become much more illuminating.

## 4. CONCLUSION

Experiments with isolated word recognition show clear, but not overwhelming, differences between the five European languages tested. The most salient language-specific features are the prevalence of homophony in French, which makes language modeling particularly important, and the increase in OOV rates for inflected languages, which makes lexical coverage an issue. We don't know whether cross-word effects would make differences among languages more pronounced. Unfortunately, our experiments with multilingual continuous speech corpora so far have resulted in such high word error rates that it is impossible to detect language-specific differences, except in vocabulary coverage. We view this set-back as temporary, however, and expect that, with sufficient training data available, technological improvements will improve performance to the point where interesting cross-linguistic differences become apparent.

## 5. REFERENCES


1. J. Barnett, P. Bamberg, M. Held, J. Huerta, L. Manganaro, and A. Weiss. "Comparative Recognition Performance in Large-Vocabulary Isolated-Word Recognition in Five European Languages". *Eurospeech '95, vol. 1*, pp. 189-192.
2. E. Eide, H. Gish, P. Jeanrenaud, and A. Mielke. "Understanding and Improving Speech Recognition Performance through the Use of Diagnostic Tools". *ICASSP 1995*, pp. 221-224.
3. Weintraub, M. "Why Are LVCSR Rates So High?" ARPA Speech Recognition Workshop, February 1996, The Arden Conference Center, Harriman, NY