TRANSCRIBING MULTILINGUAL BROADCAST NEWS USING HYPOTHESIS DRIVEN LEXICAL ADAPTATION

Petra Geutner, Michael Finke, Peter Scheytt, Alex Waibel and Howard Wactlar

Interactive Systems Laboratories
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

ABSTRACT

This paper describes first results of our DARPA-sponsored efforts toward recognizing and browsing foreign language, more specifically, Serbo-Croatian broadcast news. For Serbo-Croatian as well as many other than the most common well studied languages, the problems of broadcast quality recognition are complicated by 1.) the lack of available acoustic and language data, and 2.) the excessive vocabulary growth in heavily inflected languages that lead to unacceptable OOV-rates. We present a Serbo-Croatian large vocabulary system that achieves a 74% recognition rate, despite limited training data. Our system achieves this rate by a multipass strategy that dynamically adapts the recognition dictionary to the speech segment to be recognized by generating morphological variations (Hypothesis Driven Lexical Adaptation).

We will outline the bootstrapping and training process of the Janus Recognition Toolkit (JanusRTk) based broadcast news recognition engine: data collection, segmentation and labeling of the data according to different acoustic conditions, dictionary design, language modeling and training. The Hypothesis Driven Lexical Adaptation (HDLA) approach has been tested both on Serbo-Croatian and German news data and has achieved considerable recognition improvements. OOV-rates were reduced by 35-45%; on the Serbo-Croatian broadcast news data from 8.7% to 4.8% thereby also decreasing word error rate from 29.5% to 26%.

1. INTRODUCTION

When transcribing broadcast news data in other languages than the most common and well studied ones, problems of broadcast quality recognition are complicated by 1.) the lack of available acoustic and language data (since closed captions are typically not available), and 2.) the excessive vocabulary growth in heavily inflected languages that lead to unacceptable OOV-rates. While full-form word entries lead to excessively large vocabularies (and OOV-rates), the use of morpheme-based dictionaries also offers little relief: the combination of arbitrary morphemic affixes by way of a morphemic language model leads to an overgeneration of illegally inflected word hypotheses and thus increases error rates. This is especially the case for languages like Serbo-Croatian and German. As Serbo-Croatian is characterized by rapid vocabulary growth due to a large number of possible word inflections, we have to deal with out-of-vocabulary rates between 5 and 13%. This makes OOV-words a major source of recognition errors in multilingual broadcast news.

We present a Serbo-Croatian large vocabulary system that achieves a recognition rate of about 74%, despite very limited acoustic and language modeling training data. Our Serbo-Croatian JanusRTk based recognizer was trained on 12.5 hours of recorded speech of read newspaper articles and 27 broadcast news shows. In the following we will outline the bootstrapping and training process: data collection, segmentation and labeling of the data according to different acoustic conditions, dictionary design, language modeling and training.

Focusing on the reduction of the high OOV-rate, this work presents a two-pass recognition approach where the first pass is used to dynamically adapt the recognition dictionary to the speech segment to be recognized. The basic idea is that a large number of words in the hypothesis are recognized incorrectly because only the inflection ending is wrong, but the word-stem is recognized correctly. Often the right word was not in the dictionary thus constituting an OOV-word. By applying our multipass strategy we generate morphological variations of dictionary words only in a focused fashion (Hypothesis Driven Lexical Adaptation), thus dynamically adapting the recognition vocabulary. A second recognition run is then carried out on the adapted vocabulary.

Our approach has been tested both on Serbo-Croatian and German news data and has achieved considerable recognition improvements. In the former the OOV-rate could be decreased by 45% from 8.7% to 4.8% and in the latter case the OOV-rate dropped by 35% from 9.3% to 6.0%. Word accuracy experiments have been performed on Serbo-Croatian Broadcast News data, where we observed a relative performance improvement of 12% from 29.5% to 26% word error with an adapted vocabulary.

2. SERBO-CROATIAN BROADCAST NEWS DATABASE

Two Serbo-Croatian speech databases have been collected at the Interactive Systems Laboratory at the University of Karlsruhe: an 18 hour database consisting of read newspaper speech and a total of 18 hours of recorded and transcribed broadcast news shows.

2.1. Speech Data

The audio data for the first database, the dictation material was collected in Croatia and Bosnia-Hercegovina. Native speakers were asked to read 20 minutes of news texts extracted from the HRT (Croatian Radio and Television) web site and Obzdor Nacional, a Croatian newspaper. The speech was digitally recorded using a portable DAT-recorder at a sampling rate of 48 kHz in stereo quality and further sampled down to 16 kHz with 16 bit resolution in mono quality. The read utterances were checked against the original text to eliminate major errors and mark spontaneous effects. This data was originally collected as part of the GlobalPhone project at the University of Karlsruhe.

The broadcast news data was also collected at the University of

# Speakers	# Articles	Recording Length	# Words
85	131	18 h	89.000

Table 1: **Dictation System Database**.

Karlsruhe in Germany. A satellite dish and a dedicated PC, equipped with an MPEG encoder board, were installed to record the HRT evening news show which is transmitted from Croatia via the Eutelsat satellite. The television signal was digitally recorded in MPEG format (target bit rate: 1.008 Mbit/s, audio bit rate: 0.192Mbit/s, sampling rate: 44.1 kHz). For speech recognition the audio signal was uncompressed and sampled down to 16 kHz with 16 bit resolution. As no closed caption was available, transcription of the news broadcasts was done by native speakers. Similar to the HUB4 corpus for English broadcast news data the Serbo-Croatian recordings were divided into segments. Within these segments the acoustic conditions remained constant and each segment was tagged regarding channel quality, background noises and speaker. The various tags used in these three categories are shown in table 2, where "Non-Serbo-Croatian" identifies a person speaking in another language than Serbo-Croatian, most often English. In addition to these

Speaker	Channel	Noise
Male	Clean	Music
Female	Telephone	Second Speaker
Non-Serbo-Croatian	Distorted	Conference
Unknown	Unknown	Street
None		Static Noise
		Other
		None

Table 2: **Acoustic Segment Tags**.

acoustic tags only the most frequent and clearly audible spontaneous effects were transcribed: Hesitation, breathing and some other human and non-human noises.

It took about 13 to 18 hours to transcribe a news broadcast of approximately 40 minutes because of

- No closed caption being available,
- High speaking rate in some of the segments,
- Very noisy segments,
- Acoustic labeling of segments.

In addition to the television broadcasts, we downloaded some radio news from the Radio Free Europe/Radio Liberty web site in Realaudio format and converted them to 16 kHz, 16 bit Wave format for speech processing.

2.2. Text Data

In addition to the transcripts of the 27 news shows other sources of text data had to be collected to build up a sufficiently large corpus for language modeling purposes. Searching the internet, we retrieved

Source	Broad-casts	Recording Length	# Words
HRT (MPEG)	27	18 h	118k
RFE/RL (RA)	7	0.5 h	7k
Total	33	18.5 h	125k

Table 3: **Broadcast News System Database**.

text data from 20 different sources (television and radio stations, newspaper and news agencies). During text processing we encountered one major problem: Many sites simply map diacritics onto their corresponding non-diacritical letter, e.g. ć and č both become c. In order to build language models based on the web portion of the text corpus we had to automatically invert this mapping.

A statistical approach was used to convert web texts to usable language model training texts with diacritics. In order to get a reliable conversion as many Serbo-Croatian texts with diacritics as were available were collected. From these texts a list L_c of correct words was generated. This list served as reference to convert a second list L_f. L_f was extracted from the texts without special characters and contained both correct and false word forms. Our conversion algorithm works as follows:

- First all words in L_f that do not contain the letters c, d, s, and z are marked as correct.
- In a second step all words which occur in L_f and L_c are labeled as being correct. For some words this might be wrong in certain situations depending on the context. A word trigram model is used to improve the conversion accuracy in such cases.
- In the next step all remaining words in L_f are assigned to their nearest neighbours in L_c. When the Levenstein editing distance does not exceed a certain threshold, this word pair is considered valid and the necessary conversions are performed. When applying this operation to a separate test text, only 2% of the words were not converted correctly.
- As last step of the text conversion algorithm, we generate a letter trigram model. This model is used to score the likelihood of the different possible character sequences (switching the potential diacritic candidates c, d, s and z) for the remaining words in L_f. The sequence with the highest score is picked.

In the test text 25% of the words were converted incorrectly using this mechanism. This allows a better conversion than just leaving the words as they are, which produces an error rate of 70%. Thus finally the combined conversion error rate of the whole algorithm on the test text was 5% and enabled us to use more than twice the amount of text training material than we had before.

3. JANUS SPEECH RECOGNITION ENGINE

3.1. Dictation System

For building a Serbo-Croatian broadcast news recognizer the Janus Recognition Toolkit (JRTk) [1] was used. Phone set and a pronunciation dictionary were generated almost automatically, as Serbo-

Character Set	Web Sites	# Words
Diacritics	7	5 M
No Diacritics	13	6 M
Total	20	11 M

Table 4: **Internet Text Databases**.

Croatian orthography closely matches its pronunciation. As a consequence the phone set corresponds almost exactly to the alphabet and consists of 30 phones, 4 noise and 1 silence models. The pronunciation dictionary was created by an automatic grapheme-to-phoneme tool. Some manual adjustments were necessary for numbers, abbreviations, foreign words and names.

Each phone is modeled by a left-to-right HMM with 16 diagonal Gaussians. The preprocessing of the system consists of extracting Mel-frequency cepstral coefficients every 10 ms. The final feature vector is computed by a truncated LDA transformation of a concatenation of MFCCs and their first and second order derivatives. Vocal tract length normalization and cepstral mean subtraction are used to extenuate speaker and channel differences.

A first context-independent Serbo-Croatian dictation system was trained using the labels generated by a speaker-adapted German recognizer (label boosting [2]). The Serbo-Croatian phones were initialized by their closest German equivalents. A backoff trigram language model built on the very few available training transcriptions was used. The labels rewritten by the first Serbo-Croatian trained recognizer turned out to be more accurate and were used to train a context-dependent dictation system.

With a vocabulary size of 18k words, speaker-dependent VTLN, MLLR adaptation during testing and the use of interpolated language models from different corpora, initial system performance improved to 28.2% word error rate on the read newspaper test set using our dictation engine D1.

	Vocabulary Size	OOV-Rate	Word Error
read data	18k	8.5%	28.2%
broadcast news	18k	22.2%	73.6%

Table 5: **Initial Results for Dictation and Broadcast News System** using a Dictation System (D1).

3.2. Broadcast News System

For the broadcast news domain the test set consists of acoustic segments from two news broadcasts. Results reported below correspond to the English PE (partitioned evaluation) test set in the last HUB4 evaluation (December 1996) in which the segments and their constant acoustic properties were given for training and testing.

A first test run on the baseline system with 28.2% word error rate on the dictation data resulted in 73.6% word error rate on the broadcast news test set (see tables 5 and 6). This was mainly due to the noisy conditions even in the clean segments. The baseline dictation system was used to label our broadcast news data and train a first recognizer (B0) on only 10 hours of transcribed recordings. This context-dependent system was set up with 2k codebook vectors over 24 input features. The vocabulary size was 29k, the OOV-rate 14.0%.

System	Vocabulary Size	OOV-Rate	Word Error
D1	18k	22.2%	73.6%
B0	29k	14.0%	43.6%
B4	31k	13.6%	36.0%
B5	49k	8.7%	29.5%

Table 6: **Recognition Results** on the Broadcast News Test Set.

As interpolation of different language models resulted in performance improvements with system D1 on the dictation task, we expected the same for our broadcast news recognizer. Compared to using a single language model an absolute improvement of 1.6% word error rate was made. As we had collected text corpora from about 20 different sites, we applied three criteria to divide our text data into different sets: Geographical origin (Serbia vs. Croatia), content source (television and radio stations vs. newspaper and news agencies) and language model perplexity. An interpolation of three different text corpora yielded the best recognition results.

Further improvements were made by weighted combination of the training data of the dictation data and broadcast news data (B4). We augmented the vocabulary (31k), which slightly reduced the OOV-rate to 13.6%. The performance of the recognizer trained on those data with about twice as many parameters (mixture of gaussians) as B0 was measured to be 36% word error rate (see table 6).

Our final Serbo-Croatian broadcast news recognizer (B5) was trained on 12.5 hours dictation data and 18 hours of transcribed news shows. The context dependent system is based on 4000 quinphone models. The preprocessing of the system consists of extracting an MFCC based feature vector every 10 ms with a window size of 20 ms. The final 32-dimensional feature vector is computed by a truncated LDA transformation of a concatenation of 13 MFCCs, the energy value, their first and second order derivations, plus zero crossing. Compared to the previous B4 system, two major changes were made: 1.) text material used for language model training was normalized, 2.) the vocabulary size of the recognition dictionary was increased from 31k to 49k.

In Serbo-Croatian up to three different dialectic variations of one word can be found, e.g. for the English word 'river', the Serbian variant 'reka', but also the Croatian variants 'rjeka' and 'rijeka' exist (see table 7). When normalizing all available text material the latter two variants were replaced by the first one in all texts (language

Serbian	Croatian	Translation
reka	rjeka	river
	rijeka	

Table 7: **Serbian and Croatian Variants**.

model corpora, training and test data) and added as pronunciation variants into the dictionary. Increasing the vocabulary size from 31k to 49k led to an OOV-rate of 10.1% instead of 13.6% on the unnormalized data, and normalization further reduced the number of OOV-words to 8.7%.

Normalization	OOV-Rate
no	10.1%
yes	8.7%

Table 8: **OOV-Rates** with 49k Vocabulary.

4. VERY LARGE VOCABULARY RECOGNITION

Let N be the maximum number of words a speech recognition engine can handle in decoding. For speed and memory reasons this number is limited in current state-of-the-art recognizers to be some-

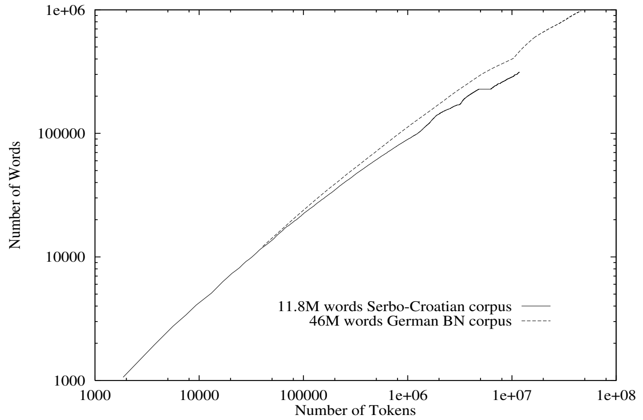

Figure 1: **Words per Token** shows the rapid vocabulary growth in German as well as Serbo-Croatian.

Word	Decomposition	Translation
žena	žen -a	woman
ženo	žen -o	woman! (vocative)
ženu	žen -u	(the) woman (accusative)
žene	žen -e	woman's (genitive)
ženi	žen -i	(to the) woman (dative)
ženom	žen -om	(with the) woman
govoriti	govor -iti	to speak
govorim	govor -im	I speak
govoriš	govor -iš	you speak (singular)
govori	govor -i	he speaks
govorimo	govor -imo	we speak
govorite	govor -ite	you speak (plural)
govore	govor -e	they speak

Table 9: **Examples for Serbo-Croatian Morphology.**

Word	Decomposition	Translation
Wahrheit	Wahr -heit	truth
Schwierigkeit	Schwierig -keit	difficulty
Kinder	Kind -er	children
Kindern	Kind -ern	children
gehen	geh -en	to go
(ich) gehe	geh -e	(I) go
(Du) gehst	geh -st	(you) go
(er) geht	geh -t	(he) goes

Table 10: **Examples for German Morphology.**

where in the range of 20k to 60k words. Constraining the maximum number of words can be considered acceptable when building recognizers for languages like English, where the number of out-of-vocabulary words given N=60k vocabulary is below one percent. With error rates for tasks like broadcast news or conversational speech (Switchboard) between 30% and 40% (due to highly disfluent speech, noisy environment, and overlapping speech, music etc.) an OOV-rate of less than a percent is not considered a major or significant source of errors[1].

As shown above for Serbo-Croatian, for languages other than English the picture is very different. In order to achieve reasonable automatic transcription performance in the broadcast news domain for languages that are characterized by rapid vocabulary growth due to a large number of possible word inflections (for Serbo-Croatian see table 9 and for German table 10), we have to expect out-of-vocabulary rates between 5% and 13%. Figure 1 shows the number of words as a function of the number of tokens in broadcast news data for both German and Serbo-Croatian. In figure 2 we compare the self-coverage and cross-coverage as measured on newspaper and broadcast news text corpora.

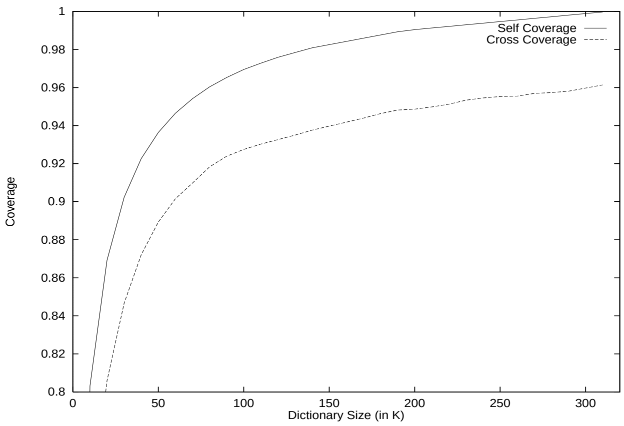

Figure 2: **Self- and Cross-Coverage** on Serbo-Croatian broadcast news data.

[1] Rule of thumb: one OOV-word causes about $1.5 - 2$ additional errors.

5. HYPOTHESIS DRIVEN LEXICAL ADAPTATION

A cheating experiment can be performed, pretending all information about the news vocabulary of a certain day would be accessible. Even when all important keywords of the day of transmission of the news broadcast would be known, the OOV-rate would only decrease to 7.8% in Serbo-Croatian news. The same cheating experiment as described above, was done on German data and resulted in an OOV-rate of 5.5%. This means that a significant portion of the OOV-words are not necessarily day or event related (new events cause new words to show up).

Therefore, the following vocabulary adaptation approach makes use of acoustic similarity instead of semantic similarity to reduce the OOV-rate. A first recognition run on a general baseline dictionary is followed by a second recognition run with a dynamically adapted dictionary of the same size but a smaller OOV-rate. Especially in a time uncritical process like the recognition of broadcast news this seems to be a practical idea.

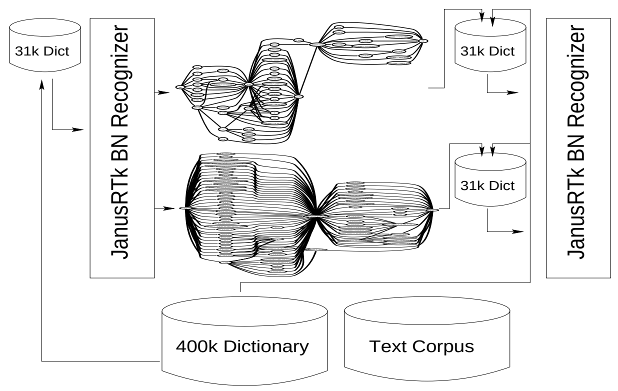

Figure 3: **Lexical Adaptation** based on Lattices. Two-pass recognition and vocabulary adaptation.

In a first recognition run, word lattices for all test utterances are created. The lattice is then used to determine, which words are most likely uttered in the segment (namely all words represented in the lattice). For each utterance to be recognized this lattice leads to an utterance-specific vocabulary. This vocabulary is then used to dynamically adapt the recognition dictionary. The basic idea is, that a large number of words in the recognized hypothesis are recognized incorrectly because only the inflection ending is wrong whereas the stem was recognized correctly. In many cases this was not due to misrecognition but because the right word was not even in the dictionary of the recognizer, so constituting an OOV-word. The algorithm below shows the whole **Hypothesis Driven Lexical Adaptation** process:

1. A first recognition run gives word lattices and an utterance-specific vocabulary list.
2. This vocabulary list is then split into word stems and suffixes (where different combinations of word stem and suffix lengths were tested, see table 11). Note that the word stem length had at least to be 2 letters long.
3. The resulting word stem list is then used to look up all similar words in the full dictionary consisting of all words that were observed in the language model training text.
4. All words with the same stem are then incorporated into the dictionary by being replaced with the least frequent words that did not show up in the lattice (so that the dictionary size of the recognizer remains N).
5. In an automatic procedure a new dictionary and language model is created to perform a second recognition run.

This vocabulary adaptation procedure applied to Serbo-Croatian broadcast news data yields a significant improvement in terms of the OOV-rate, which is reduced by 40% (see table 11), and in terms of the accuracy by reducing the error rate by 5.8% (see table 12).

	Wordstem Length				
Suffix Length	2	3	4	5	6
1	9.7%	9.0%	8.7%	8.4%	9.0%
1+2		8.9%	8.2%	8.2%	8.6%
1+2+3			8.1%	8.0%	8.4%
1+2+3+4			8.2%	7.9%	8.3%

Table 11: **Serbo-Croatian OOV-rates** with different Splitting Methods. The baseline OOV-rate is 13.6%.

	Vocabulary Size	**OOV-Rate**	**Word Error**
Baseline	31k	13.6%	36.0%
Adapted	31k	7.9%	30.2%

Table 12: **Serbo-Croatian Recognition** Results based on Adapted Vocabulary.

	Wordstem Length				
Suffix Length	2	3	4	5	6
fixed	–	–	7.7%	6.0%	6.5%

Table 13: **German OOV-rates** with different Splitting Methods. The baseline OOV-rate is 9.3%.

Table 13 shows that the same result holds for German news data, again a significant reduction of the OOV-rate. For German a fixed list of suffixes was used to create the word stems. Some examples for the used suffixes are given in table 10. Using this linguistic knowledge for decomposition also resulted in a huge OOV-rate reduction from 9.3% to 6.0% (see table 13).

In both languages it turned out to be a good choice to fix the stem length to 5 which is correlated with the distribution of word lengths (50% of the words are longer than 5 letters). Figure 4 shows the distribution of different word lengths in Serbo-Croatian and German.

	Vocabulary Size	OOV-Rate	Word Error
Baseline	49k	8.7%	29.5%
Adapted	49k	4.8%	26.0%

Table 14: **Serbo-Croatian Recognition** Results based on Adapted Vocabulary.

6. RESULTS

The same experiments as described above for system B4 were also performed on our latest B5 system. Starting off with a baseline performance of 29.5% WE and an OOV-rate of 8.7%, through HDLA we were able to reduce the number of OOV-words to 4.8%. The 3.9% improvement in OOV-rate was also reflected in a 3.5% improvement in word error rate yielding a performance of 26% WE.

7. MULTILINGUAL INFORMEDIA

The automatically generated transcripts are inserted into the multilingual Informedia database (see figure 5). In collaboration with the Informedia group at CMU [3] we have now introduced the Serbo-Croatian recognizer into a multilingual information retrieval system. Together with the Serbo-Croatian broadcast video material, the transcripts and the recognizer allow for automatic content-addressable search and multimedia document retrieval across languages (see separate system demo).

The extension of the Informedia database to more than one language will not only add to the diversity of information retrieved by monolingual queries but will also offer the possibility to phrase queries in several languages. Thus, the development of our Serbo-Croatian recognizer [4] provides an instance of a potentially larger multilingual information resource.

8. CONCLUSIONS

In this paper we described the development of a Serbo-Croatian broadcast news recognizer. It was shown that despite the very limited amount of training data a performance of 26% word error rate can be achieved. With respect to the problem of encountering excessive growth of vocabularies in heavily inflected languages like Serbo-Croatian and German, Hypothesis Driven Lexical Adaptation turned out to be a very effective means of reducing the rate of out-of-vocabulary words. By applying this two-pass recognition technique morphological variations were generated in a focused fashion which effectively reduced the number of OOV-words by 45% relative from 8.7% to 4.8% OOV-rate.

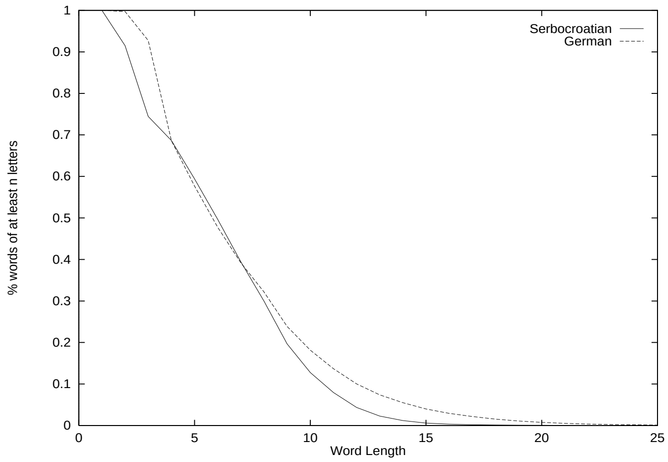

Figure 4: **Cummulative Distribution of Word Length** for German and Serbo-Croatian language.

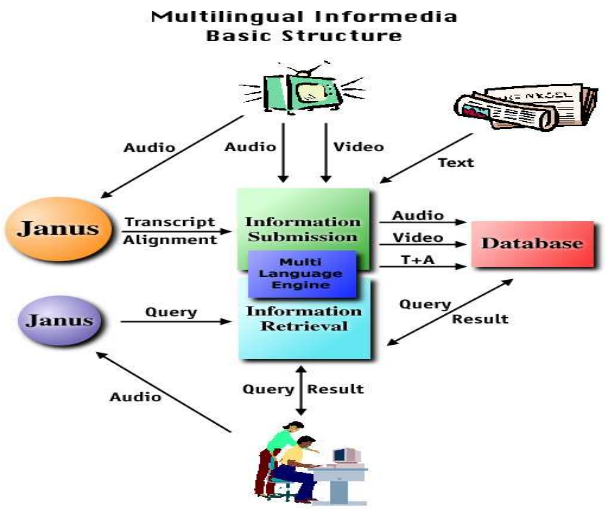

Figure 5: **Integration of the Serbo-Croatian recognizer.**

9. ACKNOWLEDGEMENTS

This research was partly funded by the Advanced Research Projects Agency under contract No. N66001-97-D-8502. The views and conclusions contained in this document are those of the authors and do not necessarily reflect the position or policy of the Government and no official endorsement should be inferred. Special thanks to Alex Hauptmann and all members of the Informedia group at Carnegie Mellon for their help and collaboration.

References

1. Finke, M., Fritsch, J., Geutner, P., Ries, K., and Waibel, A., *The JanusRTk Switchboard/Callhome 1997 Evaluation System*, Proceedings of the LVCSR Hub5-e Workshop, Baltimore, Maryland, May 1997.
2. Finke, M., and Zeppenfeld, T., *Switchboard April 1996 Evaluation Report*, Proceedings of the LVCSR Hub5-e Workshop, Baltimore, Maryland, May 1996.
3. Hauptmann, A.G., and Witbrock, M.J., *Informedia: News-on-Demand Multimedia Information Acquisition and Retrieval* in Intelligent Multimedia Information Retrieval, 1997.
4. Scheytt, P., Geutner, P., and Waibel A., *Serbo-Croatian LVCSR on the Dictation and Broadcast News Domain*, to appear in Proceedings of the IEEE 1998 International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seattle, Washington, May 1998.