Comparison of approaches for an efficient phonetic decoding

Luiza Orosanu, Denis Jouvet

Speech Group, LORIA
Inria, Villers-les-Nancy, F-54600, France
{luiza.orosanu, denis.jouvet}@loria.fr

Abstract

This article analyzes the phonetic decoding performance obtained with different choices of linguistic units. The context is to later use such an approach as a support for helping communication with deaf people, and to run it on an embedded decoder on a portable terminal, which introduces constrains on the model size. As a first step, this paper presents and analyses the performance of various approaches. Two baseline systems are considered, one relying on a large vocabulary speech recognizer, and another one relying on a phonetic n-gram language model. Then syllable-based lexicons and language models are investigated. Various lexicon sizes are studied by setting thresholds on their frequency of occurrences in the training data. Evaluations are conducted on the ESTER and ETAPE speech corpora. Keeping only the most frequent syllables leads to a limited-size lexicon and language model, which nevertheless provides good phonetic decoding performance. The phone error rate is only 4% worse (absolute) than the phone error rate obtained with the large vocabulary recognizer, and much better than the phone error rate obtained with the phone n-gram language model.

Index Terms: phonemes, syllables, deaf, speech recognition, embedded system.

1. Introduction

Support for deaf people or for people with hearing impairment is an application area of automatic speech processing technologies [1]. Their objective is to become a communication aid for disabled persons, to insure a better comprehension for the user (by the means of speech recognition) and also a better communication from the user (by the means of speech synthesis).

Over the past decades, scientists have tried to offer a better speech understanding, by displaying phonetic features to help lipreading [2], by displaying signs in sign language through an avatar [3], and of course by displaying subtitles, generated in a semi-automatic or fully automatic manner. The ergonomic aspects and the conditions for using speech recognition to help deaf people were analyzed in [4].

One of the main drawbacks of speech recognition systems is their incapacity of recognizing the words that do not belong to their vocabulary. Given the limited amount of speech training data, and also the need of a compromise between a reasonable memory use and a reasonable recognizer's performance with an acceptable execution time, it is impossible to conceive a system that covers all the words, let alone the proper names or abbreviations. When a spoken word can not be identified within the current vocabulary, the recognition system will automatically recognize it as an other one close to it, or as a series of other small words acoustically similar to the unknown word. Furthermore, recognition systems are not perfect, it happens quite frequently that a word is confused with another one which is pronounced the same (homophone) or almost the same. The performances are very far from human performance [5] and even degrade rapidly in the presence of noise. Therefore, in the context of communication aids for deaf people, displaying the orthographic form of the recognized words may not be an ideal solution.

IBM has thus tested subtitling the phonetic speech of a speaker, with the system called LIPCOM [6]. The recognition system was mono-speaker, and has been tested in a school for deaf children. The application was based on a phonetic decoding (with no prior defined vocabulary) and the result was displayed as phonemes coded on one or two letters. More recent studies have measured the contribution of confidence measures [7] within the use of automatic transcription for deaf people [8]. Subjective tests have shown a preference for displaying the phonetic form of the words with a low confidence score.

But the phoneme presents many irregularities in its phonetic realizations and a larger recognition unit, like the syllable, should help capture variations such as those introduced by coarticulation. The use of syllable-size acoustic units has been investigated in the past [9, 10, 11], for large vocabulary continuous speech recognition (usually in combination with context dependent phones) [12, 13] or for phonetic decoding only [14]. In [11], the syllables has been described as an attractive unit for recognition thanks to its greater stability, natural link between acoustics and lexical access and its capability of incorporating prosodic information into recognition. In [14], because of the structure of the acoustic units, coarticulation was modeled between phonemes inside the syllable unit, but no context-dependent modeling was taken into account between syllable units, moreover the language model applied at the syllable level was a bigram. Besides, to overcome the limited size of any speech recognizer lexicon, studies have been conducted in extending the word-based lexicon with fragments, typically sequences of phonemes determined in a data driven way; this extension helps providing better acoustic matches on out-of-vocabulary portions of the speech signal, which globally leads to a smaller phonetic error rate [15].

In this paper we shall investigate the use of syllables at the lexical level. The pronunciation of the syllables are described in terms of phonemes, which are modeled with context-dependent 3-states HMM. The language model applied on the syllables is a trigram. We have followed the rules proposed in a recent study for detecting syllables boundaries within a sequence of phonemes [16]. These rules are used to derive the syllables from the phonetic forced-aligned training data, and some criteria are applied to reduce the list of syllables constituting the lexicon. Performance is reported in terms of phoneme error rate, and evaluations are conducted on two large French speech corpus.

The work presented in this paper is part of the RAPSODIE

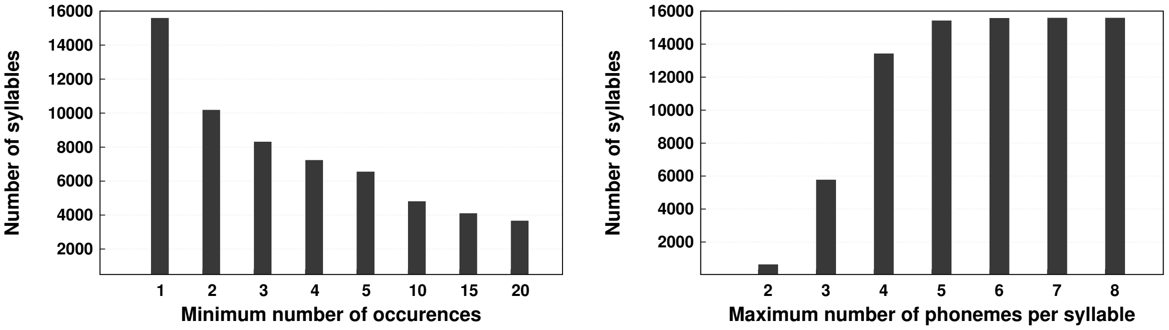

Figure 1: Number of syllables with respect to the minimum number of occurrences (left) and maximum number of phonemes per syllable (right)

project, which aims at studying, deepening and richening the extraction of relevant speech information, in order to support communication with deaf or hard of hearing people. Therefore, the optimal solution should determine the best compromise for the recognition model and the best way of presenting the recognized information (words, syllables, phonemes or combinations), within the constraints of limited available resources (the memory size and computational power of an embedded system).

This article thus performs a detailed study and analysis of the performance of various modules (in particular the choice of linguistic units and their associated language model) and heuristic decoding to obtain the best compromise between computational cost and usability results. The first section provides a description of the various linguistic units used in our analysis. The second part of the paper is devoted to the description of experiments and the discussion of results.

2. Choice of linguistic units

This section describes the linguistic units used in our analysis: the phonemes, as the basic and smallest linguistic unit, the syllables, as the phonological "building blocks" of words, and the words, as the largest linguistic unit, but at the same time the smallest linguistic element which caries a real meaning.

Note that the choice of linguistic units impacts on the choice of the vocabulary and of the language model. In the experiments reported later, the acoustic unit will always be the phoneme.

The language models used in our analysis are trigram statistical models, thus for each three lexical unit sequence, the probability of the last unit depends on the identity of the two units that precede it.

2.1. Phonemes

Regarding the pronunciation lexicon, the pronunciation of a phoneme is the phoneme itself.

Using this type of linguistic unit, we minimize the size of our vocabulary (less than 40 phonemes for the French language) and therefore the size of our language model. But unfortunately, with less modeling power usually comes worse performance.

2.2. Words

The words vocabulary contains the mappings from words to their pronunciations in the given phoneme set. Part of the reason why French pronunciation is so difficult is due to the fact that French is a non-phonetic language (some letters can be pronounced in different ways or sometimes not at all) and to "liaisons". A liaison is the phenomenon whereby a normally silent consonant at the end of a word is pronounced at the beginning of the word that follows it. So, in order to make the automatic phonetic transcription as fluid as the real speech (and thus mimic real pronunciations), scientist usually consider within the dictionary all possible "liaison" events between words. Like for example, in the sentence "les oiseaux", a classical word to phoneme transcription would give "l eh w a z o", instead of returning the real pronunciation "l eh z w a z o".

Using this type of linguistic unit leads to a large vocabulary (about 97,000 words in our dictionary) and therefore also to a large language model. This kind of model usually gives the best performance, but with the cost of great memory use and slow computational time (not ideal for embedded systems).

2.3. Syllables

Regarding the vocabulary, the pronunciation of a phonetic syllable is its decomposition into the phonemic components.

In order to account for the "liaison" events, the words will not be processed individually. The training corpora is entirely phonetized and the resulting continuous list of phonemes will be processed by the syllabification tool. The phonetization process is realized by force-aligning the manual transcriptions; each word is then replaced with its corresponding pronunciation variant found in the words vocabulary. Note that the vocabularies used in speech recognition follow real pronunciations, which means that a word can have several pronunciation variants, and that one or more phonemes might be missing in some of them.

Our syllabification tool is based on the rules described in [16], which follow two main principles: a syllable contains a single vowel and a pause designates a syllable's boundary. Therefore, the syllabification algorithm will give out models of syllables and pseudo-syllables. The pseudo-syllables are the units where one vowel is surrounded by a great number of consonants, which normally shouldn't belong to a single syllable. Using this kind of models is acceptable in automatic transcriptions , because it is quite frequent that users "skip" phonemes in their fluid, fast pronunciations ans also that is quite frequent that recognition systems "skip" too short, non-obvious phonemes. We have tried to reduce the number of abnormally-long pseudo-syllables (those that cover more than 2 words) by using as additional information the boundaries between words, but the results were not improved.

In order to filter some of the pseudo-syllable models, we have chosen to create different lists corresponding to two crite-

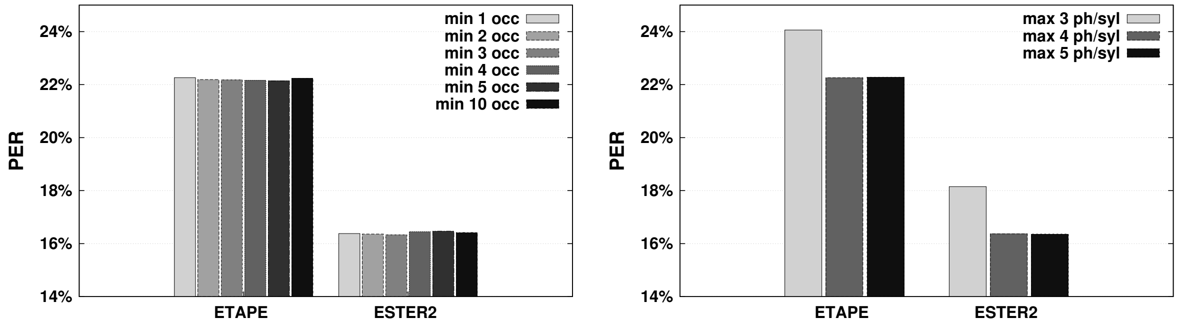

Figure 2: Performance analysis on the syllabic n-gram language models, selected according to a minimum number of occurrences (left) or to a maximum number of phonemes per syllable (right)

ria : a minimum number of occurrences within the training corpora, and a maximum number of phonemes per syllable. Figure 1 presents the number of syllables resulting from the application of each criterion (the corpora is described in section 3.1).

Using syllables as linguistic units leads to a compromise between the memory use (up to 16000 syllables within our dictionary) and computational time (ideal for embedded systems).

3. Experiments and results

This section describes the data sets and tools used in our experiments, along with the corresponding results.

3.1. Data

The speech corpora used in our experiments come from the ESTER2 [17] and the ETAPE [18] evaluation campaigns, and the EPAC [19] project. The ESTER2 and EPAC data are French broadcast news collected from various radio channels, thus they contain prepared speech, plus interviews. A large part of the speech data is of studio quality, and some parts are of telephone quality. On the opposite, the ETAPE data correspond to debates collected from various radio and TV channels. Thus this is mainly spontaneous speech.

The speech data of the ESTER2 and ETAPE train sets, as well as the transcribed data from the EPAC corpus, were used to train the acoustic models. The training data amounts to almost 300 hours of signal and almost 4 million running words. The phoneme-based language model and the syllable-based language models were also trained on the ESTER2, ETAPE and EPAC text corpora, on about 12 million running phonemes and on about 6 million running syllables.

For the creation of the word-based language model, various text corpora were used: more than 500 million words of newspaper data from 1987 to 2007; several million words from transcriptions of various radio broadcast shows; more than 800 million words from the French Gigaword corpus [20] from 1994 to 2008; plus 300 million words of web data collected in 2011 from various web sources, and thus mainly covering recent years.

For the word-based lexicon, the vocabulary of about 97,000 words, was developed for the ETAPE evaluation campaign. The pronunciation variants were extracted from the BDLEX lexicon [21] and from in-house pronunciation lexicons, when available. For the missing words, the pronunciation variants were automatically obtained using JMM-based and CRF-based Grapheme-to-Phoneme converters [22].

3.2. Configuration

The SRILM tools [23] were used to create the statistical language models. The Sphinx3 tools [24] were used to convert the SRILM language models into the Sphinx3 format and to decode the audio signals. The MFCC (Mel Frequency Cepstral Coefficients) acoustic analysis gives 12 MFCC parameters and a logarithmic energy per frame (window of 32 ms, 10 ms shift). The acoustic HMM models were modeled with a 64 Gaussian mixture, and adapted to male and female data.

3.3. Results

The development sets of the ESTER2 (non-African radios, about 42,000 running words and 142,000 running phonemes) and ETAPE (entire set, about 82,000 running words and 263,000 running phonemes) data are used in the experiments reported below.

LM	**# of *n*-grams**			**Size**
	n=1	**n=2**	**n=3**	**[MB]**
phonemes	40	1347	30898	0.21
syl_min1occ	15.6 K	0.38 M	1.74 M	10.28
syl_min2occ	10.2 K	0.38 M	1.73 M	10.07
syl_min3occ	8.3 K	0.38 M	1.73 M	9.97
syl_min4occ	7.2 K	0.37 M	1.73 M	9.90
syl_min5occ	6.5 K	0.36 M	1.73 M	9.85
syl_min10occ	4.8 K	0.35 M	1.71 M	9.65
syl_max3ph	5.8 K	0.29 M	1.56 M	8.60
syl_max4ph	13.4 K	0.38 M	1.73 M	10.14
syl_max5ph	15.4 K	0.38 M	1.74 M	10.27
words	97.3 K	43.35 M	79.30 M	1269.81

Table 1: The description of language models

The COALT (Comparing Automatic Labelling Tools) software [25] was used for the analysis of results (phoneme error rates). The compared files are the hypothesis .ctm file (resulting from the decoding process) along with the reference .stm file. The CTM file consists of a concatenation of time-marked phonemes. The STM (segment time marked) file describes the reference transcript and consists of a concatenation of text segments. We forced-aligned the STM files, in order for them to contain concatenations of time-marked phonemes as well.

Table 1 describes the language models (LM) used in our experiments. With phoneme-based units, the number of 3-grams is around 30,000 which leads to a minimum disk usage. With different lists of syllables, the number of 3-grams is around 1,700,000 which leads to an average disk usage. Using a large vocabulary, the number of 3-grams is around 79,000,000 which leads to the largest disk usages.

Below, the performance is described in terms of phoneme error rates (PER), along with the 95% confidence interval and their corresponding percentages of insertions (Ins), deletions (Del) and substitutions (Sub).

LM	PER	Ins	Del	Sub
phonemes	38.22 [±0.19]	2.87	15.41	19.94
syl_min1occ	22.26 [±0.16]	3.37	8.64	10.25
syl_min2occ	22.19 [±0.16]	3.36	8.63	10.20
syl_min3occ	22.18 [±0.16]	3.37	8.62	10.19
syl_min4occ	22.16 [±0.16]	3.36	8.62	10.18
syl_min5occ	**22.14** [±0.16]	**3.35**	**8.60**	**10.19**
syl_min10occ	22.24 [±0.16]	3.37	8.63	10.23
syl_max3ph	24.06 [±0.16]	3.63	9.34	11.10
syl_max4ph	22.26 [±0.16]	3.37	8.65	10.25
syl_max5ph	22.28 [±0.16]	3.37	8.64	10.26
words	**18.36** [±0.15]	**3.14**	**8.16**	**7.06**

Table 2: Performance analysis on ETAPE corpora [%]

Table 2 presents the results obtained on the ETAPE's development set. As expected, the best results were obtained with the large vocabulary recognizer. By using only the most frequent syllables within the language model, we limit the size of the lexicon (about 7000 syllables, cf. Figure 1) and the size of the language model (only about 10MB, cf. Table 1), and we achieve nevertheless good phonetic decoding performances. The phone error rate is only 4% worse (absolute) than the phone error rate obtained with the large vocabulary recognizer, and much better than the phone error rate obtained with the phone n-gram language model.

LM	PER	Ins	Del	Sub
phonemes	34.24 [±0.25]	3.66	11.62	18.97
syl_min1occ	16.38 [±0.19]	4.05	5.05	7.28
syl_min2occ	16.36 [±0.19]	4.05	5.05	7.26
syl_min3occ	**16.33** [±0.19]	**4.04**	**5.06**	**7.23**
syl_min4occ	16.45 [±0.19]	4.06	5.05	7.34
syl_min5occ	16.47 [±0.19]	4.04	5.06	7.36
syl_min10occ	16.42 [±0.19]	4.00	5.10	7.32
syl_max3ph	18.15 [±0.20]	4.37	5.67	8.11
syl_max4ph	16.37 [±0.19]	4.02	5.07	7.27
syl_max5ph	16.36 [±0.19]	4.04	5.04	7.28
words	**12.76** [±0.17]	**3.52**	**4.84**	**4.40**

Table 3: Performance analysis on ESTER2 corpora [%]

Table 3 presents the results obtained on ESTER2's development set. We notice the same performance behavior as for the ETAPE corpora: best results for the large vocabulary recognizer, slight decrease (4% absolute) for the syllabic n-gram language models and worst results for the phone n-gram language model.

Figure 2 displays the results obtained on different syllabic n-gram language models, by exploiting the filters resulting from a minimum number of occurrences or from a maximum number of phonemes per syllable. The worst performance is obtained on the list of syllables with maximum 3 phonemes (less than 6000 models of syllables and pseudo-syllables). Besides that, all the other filters give more or less the same results. Which means that starting with a minimum number of 7,000 linguistic units we can achieve similar results as with the total number of ∼16,000 units.

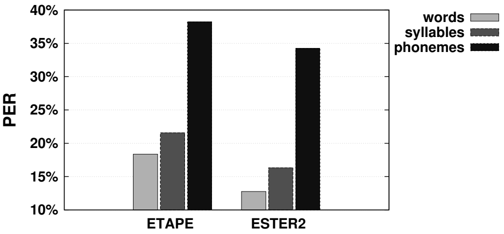

Figure 3: Summary of the results

Figure 3 displays a summary of the results obtained on both corpora. Given that ESTER2 contains mainly prepared speech and that ETAPE contains mainly spontaneous speech, the results obtained on ESTER2 are, as expected, better than the ones obtained on ETAPE.

4. Conclusions

This paper presented a detailed study on the phonetic decoding performance of various modules on two French speech corpora (ETAPE and ESTER2). We were interested in finding the best compromise between computational cost and usability of results, constrains that must be met in order to be able to create an embedded speech recognition decoder on a portable terminal. The context is to later use such an approach as a support for helping communication with deaf people.

Two baseline systems were considered. The first one relies on a large vocabulary speech recognizer; it gives the best results (∼18% phoneme error rate (PER) on ETAPE and ∼12% PER on ESTER2), but it uses a lot of memory and computational power. The second one relies on a phonetic n-gram language model; it does not use much memory, nor computational power, but it does not give good results neither (∼38% PER on ETAPE and ∼34% PER on ESTER2).

Then syllable-based lexicons and associated 3-gram language models were investigated. The lexicons of syllables and pseudo-syllables were filtered according to the number of occurrences in the training data and the number of phonemes. Keeping only the most frequent syllables leads to a limited-size lexicon and language model, which nevertheless provides good phonetic decoding performance. The phone error rate is only 4% worse (absolute) than the phone error rate obtained with the large vocabulary recognizer, and much better than the phone error rate obtained with the phone n-gram language model.

Future work will focus on the best, suitable way of presenting the recognized information (phonemes, syllables, words or combinations), based on relevant confidence measures, so that it maximizes communication efficiency with deaf people.

5. Acknowledgements

The work presented in this article is part of the RAPSODIE project (http://erocca.com/rapsodie), and has received support from the "Conseil Régional de Lorraine" and from the "Région Lorraine" (FEDER).

6. References

[1] Schönbächler, J., “Le traitement de la parole pour les personnes handicapées”, travail de séminaire, Fribourg, 2003.

[2] Sokol, R., “Réseaux neuro-flous et reconnaissance de traits phonétiques pour l’aide à la lecture labiale”, Thèse Université de Rennes, 1996.

[3] Cox, S., Lincoln, M., Tryggvason, J., Nakisa, M., Wells, M., Tutt, M. and Abbott, S., “Tessa, a system to aid communication with deaf people”, Proc. Assets’02, Vth Int. ACM Conf. on Assistive Technologies, pp. 205-212, 2002.

[4] Woodcock, K., “Ergonomics and automatic speech recognition applications for deaf and hard-of-hearing users”, Technology and Disability, vol. 7, pp. 147-164, 1997.

[5] Lippmann, R., “Speech recognition by machines and humans”, Speech Communication, n° 22, pp. 1-15, 1997.

[6] Coursant-Moreau, A. and Destombes, F., “LIPCOM, prototype d’aide automatique à la réception de la parole par les personnes sourdes”, Glossa, n° 68, pp. 36-40, 1999.

[7] Jiang, H., “Confidence measures for speech recognition: A survey”, Speech Communication, vol. 45, n° 4, pp. 455-470, 2005.

[8] Razik, J., Mella, O., Fohr, D. and Haton, J.-P., “Transcription automatique pour malentendants : amélioration à l’aide de mesures de confiance locales”, Journées d’Etude de la parole, 2008.

[9] Zhang, L. and Edmondson, W. H., “Speech recognition using syllable patterns”, 7th International Conference on Spoken Language Processing, 2002.

[10] Tachbelie, M., Besacier, L. and Rossato, S., “Comparison of syllable and triphone based speech recognition for Amharic”, Proceedings of the LTC 2011, pp. 207–211, 2011.

[11] Wu, S.-L., Kingsbury, B., Morgan, N. and Greenberg, S., “Incorporating Information from Syllable-length Time Scales into Automatic Speech Recognition”, ICASSP, pp. 721-724, 1998.

[12] Ganapathiraju, A., Hamaker, J., Ordowski, M., Doddington, G. and Picone, J., “Syllable-based large vocabulary continuous speech recognition”, IEEE Transactions on Speech and Audio Processing., vol. 9, no. 4, pp. 358–366, 2001.

[13] Hämäläinen, A., Boves, L. and de Veth, J., “Syllable-Length Acoustic Units in Large-Vocabulary Continuous Speech Recognition”, Proceedings of SPECOM, 2005.

[14] Blouch, O., Collen, P., “Reconnaissance automatique de phonemes guide par les syllables”, Journées d’Etude de la parole, 2006.

[15] Rastrow, A., Sethy, A., Ramabhadran, B. and Jelinek, F., ”Towards using hybrid, word, and fragment units for vocabulary independent LVCSR systems”, Proceedings Interspeech’2009, 2009.

[16] Bigi, B., Meunier, C., Bertrand, R. and Nesterenko, I. “Annotation automatique en syllabes d’un dialogue oral spontané”, Journées d’Etude de la parole, 2010.

[17] Galliano, S., Gravier, G. and Chaubard, L., “The ESTER 2 evaluation campaign for rich transcription of French broadcasts”, Proceedings INTERSPEECH’2009, 2009.

[18] Gravier, G., Adda, G., Paulson, N., Carre, M., Giraudel, A. and Galibert, O., “The ETAPE corpus for the evaluation of speech-based TV content processing in the French language”, 8th International Conference on Language Resources, Evaluation and Corpora, 2012.

[19] Estève, Y., Bazillon, T., Antoine, J., Béchet, F. and Farinas, J., “The EPAC corpus: Manual and automatic annotations of conversational speech in French broadcast news”, in Proc. LREC’2010, European conference on Language Resources and Evaluation, 2010.

[20] Mendonça, Â., Graff, D., DiPersio, D., “French gigaword third edition”, Linguistic Data Consortium, 2011.

[21] M. de Calmès, and G. Pérennou, “BDLEX : a Lexicon for Spoken and Written French”, Language Resources and Evaluation, pp.1129-1136, 1998.

[22] Illina, I., Fohr, D. and Jouvet, D., “Grapheme-to-Phoneme Conversion using Conditional Random Fields”, Proceedings INTERSPEECH’2011, 2011.

[23] Stolcke, A., “SRILM an Extensible Language Modeling Toolkit”, 7th International Conference on Spoken Language Processing, 2002.

[24] Placeway, P., Chen, S., Eskenazi, M., Jain, U., Parikh, V., Raj, B., Ravishankar, M., Rosenfeld, R., Seymore, K., Siegler, M., Stern, R. and Thayer, E., “The 1996 Hub-4 Sphinx-3 System”, Carnegie Mellon University, 1996.

[25] Fohr, D. and Mella, O., “CoALT: A Software for Comparing Automatic Labelling Tools”, Language Resources and Evaluation, 2012.