Internal Use Only（非公開）

TR－SLT－0086

A method to quantify corpus similarity and its application to quantifying the degree of literality in a document

Etienne DENOUAL

2004年10月21日

概要

Comparing and quantifying corpora is a key issue in corpus based translation and corpus linguistics, for which there is still a notable lack of measures. This makes it difficult for a user to isolate, transpose, or extend the interesting features of a corpus to other NLP systems. In this work we address the issue of measuring similarity between corpora. We suggest a scale between two user chosen corpora on which any third given corpus can be assigned a coefficient of similarity, based on the cross-entropy of statistical N-gram character models. A possible application of this framework is to quantify similarity in terms of literality (or conversely, orality). To this end we carry out experiments on several well-known corpora in both English and Japanese language, and show that the defined similarity coefficient is robust in terms of language and model order variations. Within this framework we further investigate the notion of homogeneity in the case of a large multilingual resource.

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619－0288「けいはんな学研都市」光台二丁目2番地2 TEL：0774－95－1301

Advanced Telecommunications Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai "Keihanna Science City" 619-0288,Japan
Telephone:+81-774-95-1301
Fax :+81-774-95-1308

A method to quantify corpus similarity *and its application to quantifying the degree of literality in a document*

Etienne Denoual
ATR - Spoken Language Translation Research Labs, Kyoto, Japan
CLIPS - GETA - IMAG, Univ. Joseph Fourier, Grenoble, France

Abstract

Comparing and quantifying corpora is a key issue in corpus based translation and corpus linguistics, for which there is still a notable lack of measures. This makes it difficult for a user to isolate, transpose, or extend the interesting features of a corpus to other NLP systems. In this work we address the issue of measuring similarity between corpora. We suggest a scale between two user chosen corpora on which any third given corpus can be assigned a coefficient of similarity, based on the cross-entropy of statistical N-gram character models. A possible application of this framework is to quantify similarity in terms of literality (or conversely, orality). To this end we carry out experiments on several well-known corpora in both English and Japanese language, and show that the defined similarity coefficient is robust in terms of language and model order variations. Whithin this framework we further investigate the notion of homogeneity in the case of a large multilingual ressource.

1 Introduction

Be it in corpus linguistics or data-driven automatic translation, statistical or example-based, corpora not only are useful tools but key elements of the discipline. Paradoxically little work has been issued on automatically characterising such sets, and attempting a meaningful comparison is often a perilous task. Typically we will read that such corpus is made of "casual speech transcripts" or "mildly spontaneous utterances", such other of "highly scenarised oral language". However such tags will be of little use to the user seeking to better understand how his system performs, and how to transpose its specific features to another task.

In the following work we try to fill this void with a method to position a corpus relatively to two others, which we use as references. As an experiment we apply this method in particular to measuring the literality (and conversely, orality) of a corpus. Firstly we examine the notion of corpus similarity, then we suggest an information theory based measure of similarity using statistical N-gram character models. Our

framework is then tested on the particular case of evaluating the literality of several corpora.

2 Corpus similarity

We are regularly provided with a wider range of corpora to use as tools in analysis or machine translation, and while we get more data everyday, it becomes harder to quickly grasp the nature of what we study. [Kilgarriff and Rose 1998, Kilgarriff 2001] investigate on the similarity and homogeneity of corpora and pinpoint the need to find appropriate measures for them, without which it is difficult to discuss the relevance of one's findings or port them to another domain. They then proceed to compare "Known Similarity Corpora" (KSC) using perplexity and cross-entropy on words, word frequency measures, and a χ^2-test which they find to be the most robust. However, as acknowledged in [Kilgarriff and Rose 1998], using KSC requires that the two corpora chosen for comparison are sufficiently similar that the most frequent lexemes in them almost perfectly overlap. Whereas intuition would hint at this being true for very large corpora, [Liebscher 2003] shows by comparing frequency counts of different Google Group corpora that it is not the case. Furthermore, while this measure gives us an idea of the similarity of such corpus to such other, it does not rank several corpora in similarity, i.e. it gives us no idea of the "distance" in similarity between them. Measuring similarity by counting word/lexeme frequencies introduces further difficulties: this assumes that the word is an immediate, well-defined unit, which is not the case in the Chinese([Sproat and Emerson 2003]) or Japanese language ([Matsumoto et al., 2002]) for instance, where word segmentation is still an unsolved issue.

How do two corpora relate to each other? Perhaps it would be easier and more intuitive to answer the question: if I define a scale between two reference corpora, where in between does this third corpus fit ? [Biber 1988, Biber 1995] identifies a set of seven dimensions by counting linguistic features in text samples, and shows that a document of text can be assigned a score on any dimension. We will use this property along with information theory to define a scale of similarity between two corpora on which any given third corpus can be assigned a similarity coefficient, with no need of prior linguistic feature selection.

3 Quantifying similarity

3.1 N-gram cross-entropy

Entropic measures performed on characters have the obvious advantage of being blindly applicable to any electronic data, without the use of any prior linguistic knowledge (thus eliminating the bias of segmentation errors in languages where the word unit is not clearly defined). In this study we turn our attention to cross-entropy in terms of N-grams of characters. [Dunning 1994] showed the interest of character based models for language identification in the way that the required training and test sets are surprisingly small in order to achieve good results, achieving a 99.9% accuracy in

identifying English from Spanish with only 50K bytes of training and 500 bytes of test text. He further shows that the accuracy of classification only improves with longer test data. The cross-entropy $H_T(A)$ of an N-gram model p constructed on a training corpus T, on a test corpus $A = \{s_1, .., s_Q\}$ of Q sentences with $s_i = \{c_1^i..c_{|s_i|}^i\}$ a sentence of $|s_i|$ characters is:

$$H_T(A) = \frac{\sum_{i=1}^{Q}[\sum_{j=1}^{|s_i|} -logp_j^i]}{\sum_{i=1}^{Q}|s_i|} \tag{1}$$

where $p_j^i = p(c_j^i|c_{j-N+1}^i..c_{j-1}^i)$. The recurrent idea in this work is to construct several N-gram character models using reference training corpora and then use these language models to estimate the cross-entropy in terms of bits per character (shortened as bpc) needed to encode a test corpus.

3.2 A coefficient of similarity

3.2.1 Definition

More than just categorising or clustering corpora, we wish to quantify similarity. We therefore define a scale of similarity between two corpora on which to rank any third given one. This is done by letting the user select two corpora T_1 and T_2, which he choose to be his references, and use them as training sets to compute N-gram character models. We estimate their cross-entropies on a third test set T_3, which we will respectively name $H_{T_1}(T_3)$ and $H_{T_2}(T_3)$ according to the notation in Equ. 1. We also estimate both cross-entropies of each reference for each model according to the other one, i.e. $H_{T_1}(T_2)$ and $H_{T_1}(T_1)$, $H_{T_2}(T_1)$ and $H_{T_2}(T_2)$ so as to obtain the weights W_1 and W_2 of references T_1 and T_2 :

$$W_1 = \frac{H_{T_1}(T_3) - H_{T_1}(T_1)}{H_{T_1}(T_2) - H_{T_1}(T_1)} \tag{2}$$

and :

$$W_2 = \frac{H_{T_2}(T_3) - H_{T_2}(T_2)}{H_{T_2}(T_1) - H_{T_2}(T_2)} \tag{3}$$

After which we assume W_1 and W_2 to be the weights of the barycentre between our chosen references. Thus, we define :

$$I(T_3) = \frac{W_1}{W_1 + W_2} = \frac{1}{1 + \frac{W_2}{W_1}} \tag{4}$$

to be the similarity coefficient between reference sets one and two, respectively corpus T_1 and corpus T_2 .

3.2.2 Meaning

Given the previous assumptions, $I(T_1) = 0$ and $I(T_1) = 0$; furthermore, any given corpus T_3 will be then awarded a score between $I(T_1) = 0$ and $I(T_2) = 1$. Here we consider two corpora similar when one of them can be completely predicted given the knowledge of the other one (i.e. given a language model constructed on the other one).

This idea is extended to three corpora, two of them being references, the third one being studied.

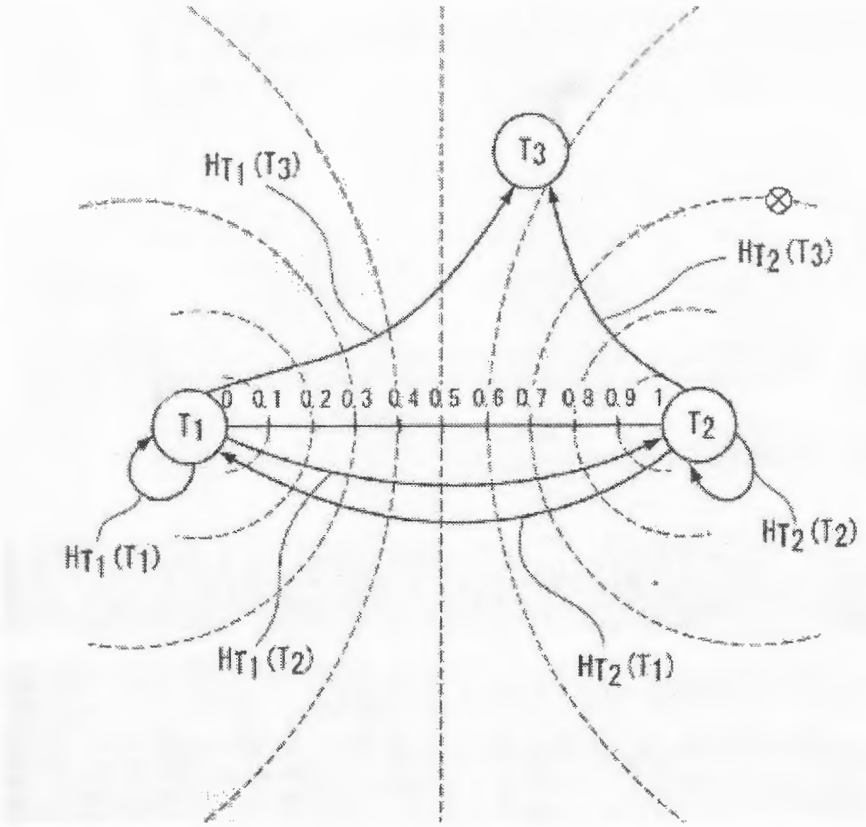

Figure 1: Method overview. Dashed lines show coefficient equivalues. (0 is oral, 1 is literal) The point $\otimes$ will be explained in section 4.3.

3.2.3 Degenerate cases

As can be seen on the schematic representation on figure 1, $I(T_3)$ may take but three values $(0, \frac{1}{2}, 1)$ when the size $n-1$ of the models history length gets close to infinity. Let B be the standard value in bits needed to encode a character in a chosen coding system, then:

- If T_3 tends be to most similar to T_2 and most dissimilar to T_1, i.e. if with a long history length $n-1$, T_3 is be completely determined by a language model trained on T_2, but is completely undetermined by a language model trained on T_1, then:

$$\lim_{n\to\infty} H_{T_1}(T_3) = B \qquad \lim_{n\to\infty} H_{T_2}(T_3) = 0 \tag{5}$$

T_1 and T_2 are assumed to be dissimilar, therefore:

$$\lim_{n\to\infty} H_{T_2}(T_1) = \lim_{n\to\infty} H_{T_1}(T_2) = B \tag{6}$$

In the same manner, a model used to predict the very corpus it was constructed upon will yield a null uncertainty for long history lengths:

$$\lim_{n\to\infty} H_{T_1}(T_1) = \lim_{n\to\infty} H_{T_2}(T_2) = 0 \tag{7}$$

We can therefore simplify the expressions of W_1 and W_2:

$$\lim_{n\to\infty} W_1 = \lim_{n\to\infty} \frac{H_{T_1}(T_3) - H_{T_1}(T_1)}{H_{T_1}(T_2) - H_{T_1}(T_1)} = \frac{B-0}{B-0} = \frac{B}{B} = 1 \tag{8}$$

$$\lim_{n\to\infty} W_2 = \lim_{n\to\infty} \frac{H_{T_2}(T_3) - H_{T_2}(T_2)}{H_{T_2}(T_1) - H_{T_2}(T_2)} = \frac{0-0}{B-0} = 0 \tag{9}$$

Consequently, if the size $n-1$ of the history length gets close to infinity then:

$$\lim_{n\to\infty} I(T_3) = 1 \tag{10}$$

- Conversely, if T_3 tends to be most similar to T_1 and most dissimilar to T_2, i.e. if with a long history length $n-1$, T_3 can be completely determined by a language model trained on T_1, but is completely undetermined by a language model trained on T_2, then:

$$\lim_{n\to\infty} W_1 = 0 \qquad \lim_{n\to\infty} W_2 = 1 \qquad \lim_{n\to\infty} I(T_3) = 0 \tag{11}$$

- If T_3 tends to be as much similar to T_2 as it is to T_1:

$$\lim_{n\to\infty} H_{T_1}(T_3) = \lim_{n\to\infty} H_{T_2}(T_3) = H \tag{12}$$

Therefore:

$$\lim_{n\to\infty} W_1 = \frac{H}{B} = \lim_{n\to\infty} W_2 \tag{13}$$

$$\lim_{n\to\infty} I(T_3) = \frac{1}{1+\frac{H/B}{H/B}} = \frac{1}{1+1} = \frac{1}{2} \tag{14}$$

We have shown here that when the size $n-1$ of the history length gets close to infinity, then I may take only one of the three values $(0, \frac{1}{2}, 1)$.

3.3 A word on corpus homogeneity

Comparing non-homogeneous corpora, for instance concatenated texts of various sources or corpora of different sizes could seem vain, for to this date we do not know of a satisfying definition of corpus homogeneity, nor of any influence of homogeneity on computation. Intuitively however, a corpus made of a collection of software manuals should be more homogeneous than one gathering mixed sentences from both telephone transcriptions and litterature pieces [Kilgarriff and Rose 1998].

In our first approach, in order to bypass this problem, we first ensure that our corpora individually originate from a same background (i.e. one is a collection of newspaper articles only, another is a collection of telephone trancripts only). We thus divide each reference corpus in n randomly selected blocs of equal size, compute cross-entropies using these blocs as test-sets, and average the results (a process usually referred to as "n-fold cross-validating", which ensures that results are not artifacts of accidentally selecting unrepresentative testing data [Charniak 1993]).

In our second experiment, we will specifically address the issue of homogeneity in the case of a large multilingual ressource.

4 Quantifying literality among corpora

4.1 Training and Test Data

An experiment was carried out on both English and Japanese language, with the hope of distinguishing and ranking similarity irrespective of the language used. In order to validate the previously described framework, we chose to set up a scale of literality between two corpora of contrasting origins:

- As a reference for "orality" we used for both English and Japanese language the SLDB (Spontaneous Speech Database) corpus, a multilingual corpus of raw transcripts of dialogues described in [Nakamura et al., 1996].
- As a reference for "literality" for the English language we used a part of the Calgary[1] corpus, familiar in the data-compression field, containing several classical and contemporary English literature pieces[2], and for the Japanese language a corpus of collected articles from the Nikkei Shinbun newspaper[3].

Test data was chosen in both English and Japanese with the hope of measuring the literality of two corpora originating from the same domain, but having a hearsay reputation of differing slightly in terms of orality. Those two multilingual corpora are the C-STAR[4] part of an aligned multilingual corpus, the Basic Traveller's Expressions Corpus (BTEC[5]), and the Machine-translation-Aided bilingual spoken Dialogue corpus (MAD), both collections of sentences from the tourism and travel domain, MAD being a collection of realistic but clean transcripts of dialogues and BTEC being a collection of sentences from travel handbooks. MAD has the reputation of being more "oral" than the BTEC. This is precisely what we wanted to measure.

[1]The Calgary Corpus is available via anonymous ftp at ftp.cpcs.ucalgary.ca/pub/projects/text.compression.corpus .

[2]Parts are entitled book1, book2 and book3.

[3]The use of classical Japanese literature is not appropriate as (older) copyright free works make use of a considerably different language. In order to maintain a certain homogeneity, we limit our study to contemporary language.

[4]See http://www.c-star.org .

[5]A summary of the abbreviations used in this paper to refer to the different corpora, along with typical utterances and any further information can be found in the Appendix.

To further validate our measure, we also wished to test it on other corpora from specific backgrounds:

- For the English language, the TIME[6] corpus (a collection of newspaper articles of the TIME magazine from the nineteen-sixties), and the corpus of Spoken Professional American-English[7], (SPAE) a collection of transcripts of meetings and interactions in professional settings.
- For the Japanese language, a corpus of collected articles from the Mainichi Shinbun newspaper (MAINICHI), and a corpus of clean transcripts of broadcasts from the NHK news.

In the following section we outline for each corpus basic statistical figures, then compute cross-entropies and the derived literality coefficient.

4.2 Statistical aspects

English corpora	SLDB	MAD	BTEC	SPAE	TIME	Calgary
Word/Sent.(Mean)	11.27	9.29	5.94	23.34	23.17	20.21
Word/Sent.(Std dev.)	6.85	5.83	3.25	26.43	15.32	15.18
Char./Sent.(Mean)	64.51	44.86	31.15	126.11	131.74	107.70
Char./Sent.(Std dev.)	35.95	27.57	17.02	140.71	92.38	84.69
Char./Word	5.72	4.83	5.24	5.40	5.68	5.33
Total Char.	1,037K	475K	5,026K	223K	1,515K	757K
Total Words	181.2K	98.5K	964.2K	41K	264.5K	142.2K
Total Sent.	16,078	10,601	162,318	1,759	11,416	7,035

Figure 2: Statistical aspects of several English corpora.

Japanese corpora	SLDB	MAD	BTEC	NHK	Mainichi	Nikkei
Char./Stce (Mean)	32.61	26.87	14.45	65.39	37.73	44.21
Char./Stce (Std dev.)	22.22	14.07	7.12	39.16	31.88	28.34
Total Char.	20,806K	290K	2,426K	2,772K	2,740K	2,772K
Total Sent.	84,751	10,612	162,318	66,512	71,647	253,016

Figure 3: Statistical aspects of several Japanese corpora.

[6]See http://www.time.com .

[7]See http://www.athel.com/cspa.html .

Statistical aspects for each corpus are shown on Figure 2 and 3 for English and Japanese. Three corpora, namely SPAE, TIME and the reference for written expression, Calgary, have notably superior Words/Sentence ratios than the three others, namely MAD, BTEC and the reference for oral expression, SLDB: the former, that we assumed more literal are around 20 words per sentence, whereas the latter that we conversely assumed to be more oral are closer to 10. This follows the general belief that written sentences tend to be longer than oral ones. However, characters per word scores are comparable for all corpora and therefore do not show any noticeable difference between a set we priorly assumed to be more oral and another one to be more literal, but are rather typical of the English language as a whole. Such statistical figures therefore do not provide enough information to differenciate between oral or written documents reliably.

Word segmentation ambiguity in the Japanese language has been the subject of intense research (for instance [Matsumoto et al., 2002]), and is a complex issue. Therefore we prefered to focus on character counts here. As well as with English corpora, characters per sentence scores for Japanese language corpora that we assumed more literal (NHK, Mainichi and Nikkei) are usually higher than the ones we priorly assumed to have a more oral content. However, the character per sentence score for SLDB (oral reference) is comparable to the one of the Mainichi, making it difficult once again to reach a reliable classification.

4.3 Entropy

Language models for N-grams ranging from $N = 2$ to 16 were computed for the two references in each language. Cross-Entropy is then computed and averaged on randomly selected, non-overlapping blocks of approximately $250,000$ characters for each corpus. Results for the English language are shown on Figure 4, for the Japanese language on Figure 5.

While N-gram orders of 4 to 6 generally achieve the lowest bits per character ratios (apart from the cross-entropy computed on the training set, for which prediction performance is optimal), all numbers rise and stagnate as the order increases. This is due to the fact that for higher orders, unrecognised N-grams increase exponentially and lead to the familiar problem of training data sparseness.Drawing from the field of data-compression where history length is an active area of research, we will assume as in [Teahan and Cleary 1997] and [Dunning 1994] that a history of 3 to 7 characters gives meaningful results.

Meeting the intuition, the cross-entropy of SLDB against itself is the minimum (the accuracy of a language model predicting the corpus it was constructed on cannot be beat).

It is not necessary for the other reference to yield the highest values. For exemple on the left of Figure 4, TIME yields the highest values and will therefore be in the situation of point $\otimes$ in Figure 1. Its projection will be within the interval [SLDB,CALGARY].

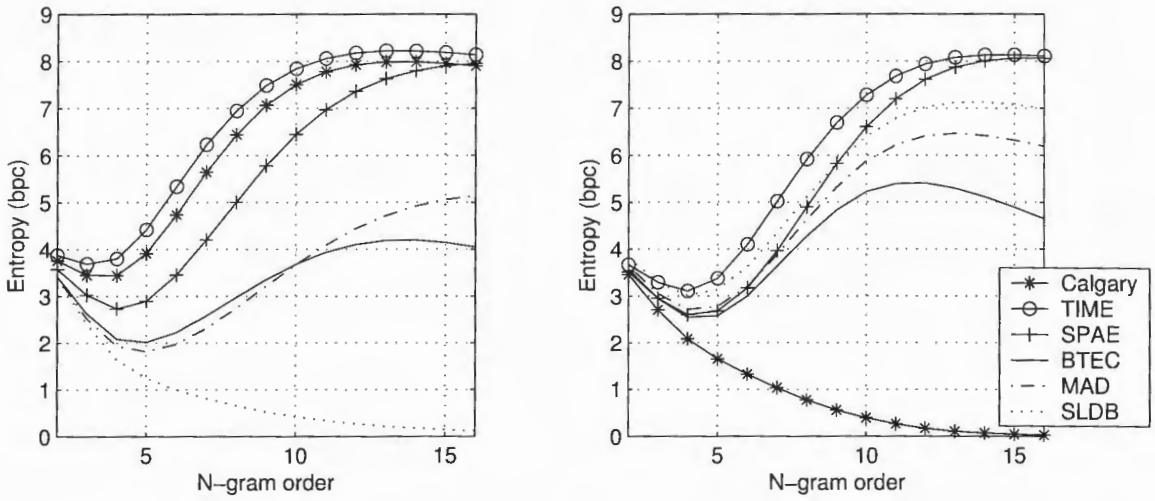

Figure 4: Cross-entropy against the oral reference SLDB (on the left) and the literal reference Calgary (on the right) respectively, for the English language.

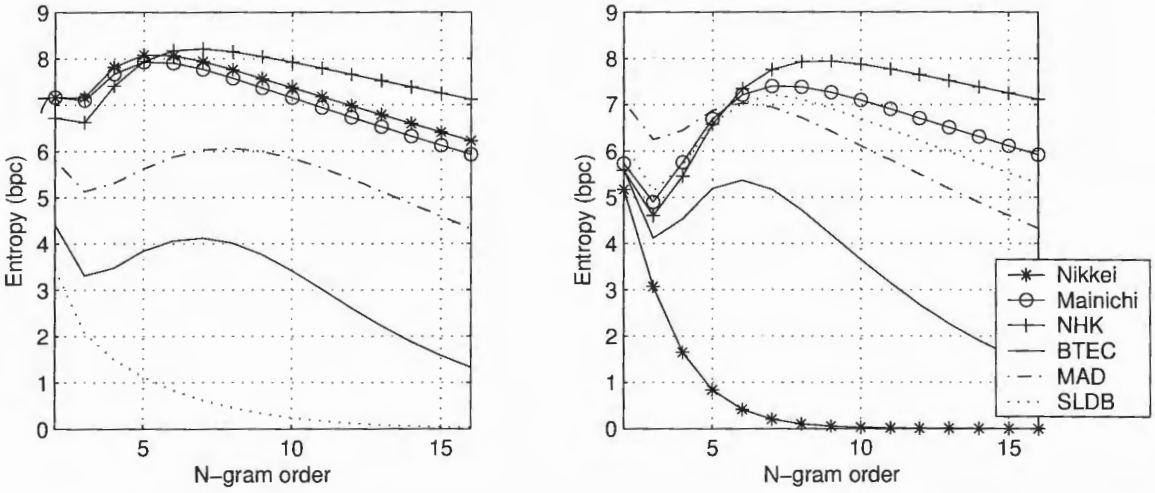

Figure 5: Cross-entropy against the oral reference SLDB (on the left) and the literal reference Nikkei (on the right) respectively, for the Japanese language.

4.4 A literality coefficient

We then compute the literality coefficient for each corpus : Figures 6 and 7 show the variations of this coefficient for different N-gram orders, respectively for English and Japanese language. According to our choice of references, a score of 0 corresponds to being closely similar to the oral reference, the SLDB corpus, whereas a score of 1 to being closely similar to the literal reference, the Calgary book corpus for English, or the Nikkei newspaper corpus for Japanese.

Figure 8 shows values for 5-gram character models.

Both MAD and BTEC yield lower literality scores than TIME and SPAE, and NHK and Mainichi respectively in English and Japanese. MAD indeed yields the lowest

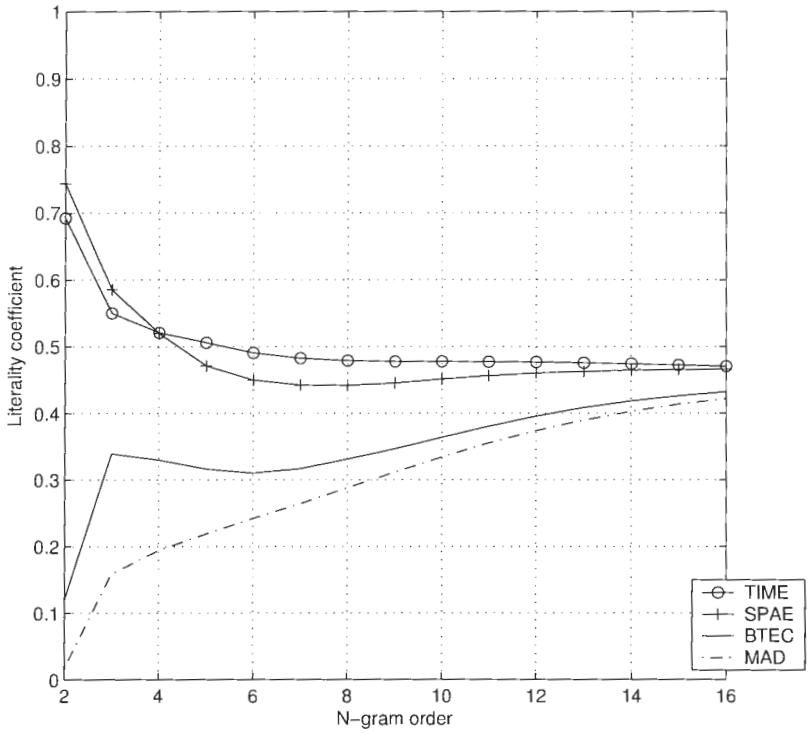

Figure 6: Literality coefficient for the English language. (0 is oral as in SLDB, 1 is literal as in Calgary)

literality scores for the English language, which would tend to confirm its reputation of being slightly more oral in content than the BTEC, both being usually refered to as "oral corpora".

It is not that obvious for the Japanese language, where the figure is less stretched out and we are hard put to reach a conclusion on literality.

For the English language, the TIME and SPAE have close values for orders of 3 to 5. However, the TIME seems to have a slightly higher literality score for all N-gram orders superior to 4. We may assume that indeed, the TIME's degree of literality is superior to SPAE's, which should intuitively be true considering that the TIME is made of written journalistic texts, whereas SPAE is made of transcripts of formal, professional conversations.

For the Japanese language, while differences in literality are clear between dialogue transcript corpora such as MAD and BTEC, and news corpora such as NHK and Mainichi, we are hard put to reach any decision at all when it comes to differenciating transcripts of highly scenarised, redactional content (NHK) from pure newspaper articles (Mainichi). We assume that the scale contraction phenomenon in the case of the Japanese language (and conversely scale stretch in the case of the English language) is due to the fact that we are unable to use a corpus of contemporary literature in the Japanese language. Classification seems robust to N-gram order variations, and all indexes converge to values between 0.4 to 0.5 for N superior to 14 for English, and as

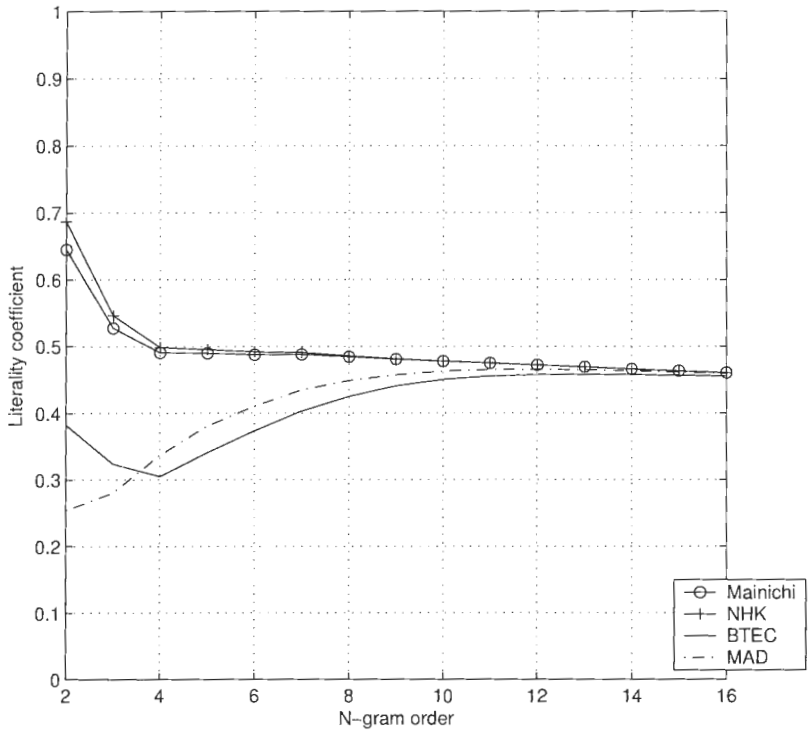

Figure 7: Literality coefficient for the Japanese language. (0 is oral as in SLDB, 1 is literal as in Nikkei)

	MAD	BTEC	SPAE/ NHK	TIME/ Mainichi
Eng	0.22	0.32	0.47	0.51
Jpn	0.38	0.34	0.49	0.49

Figure 8: Coefficient of literality for 5-gram character models.

soon as N exceeds 7 for Japanese, according to our discussion about degenerate cases in Section 3.2.3.

4.5 Measuring homogeneity inside a large multilingual ressource

Having caracterised a corpus among others by measuring similarity, we would like to profile a single corpus more particularly. Indeed, if similarity is an inter-corpus measure, then intuitively homogeneity should be an intra-corpus one. We want to apply our literality coefficient to subsets of a same corpus to try to caracterise its homogeneity in literality. As such, homogeneity will be caracterised by the internal variations of the previously defined similarity coefficient.

In this second experiment, we specifically address the issue of homogeneity whithin a large multilingual ressource, the Basic Traveller's Expressions Corpus .

To this end, the ressource was cut into a number of subsets, each of which was scored on the previously defined literality coefficient, using the same reference corpora as in the previous experiment (SLDB for orality and Calgary or Nikkei for literality). As we increase the number of subsets (anologically to increasing smoothing resolution), we aim at having a better idea of the local variations of the coefficient, and of the ideal trade-off between a high number of small subsets (less smoothed therefore more detailed, less data therefore less significant) and a low number of large subsets (more smoothed and therefore less detailed, more data therefore more significant).

The experiment is conducted on the aligned BTEC corpus in both English and Japanese language. Figure 9 shows the variation of the coefficient for 10 and 100 subsets in both languages.

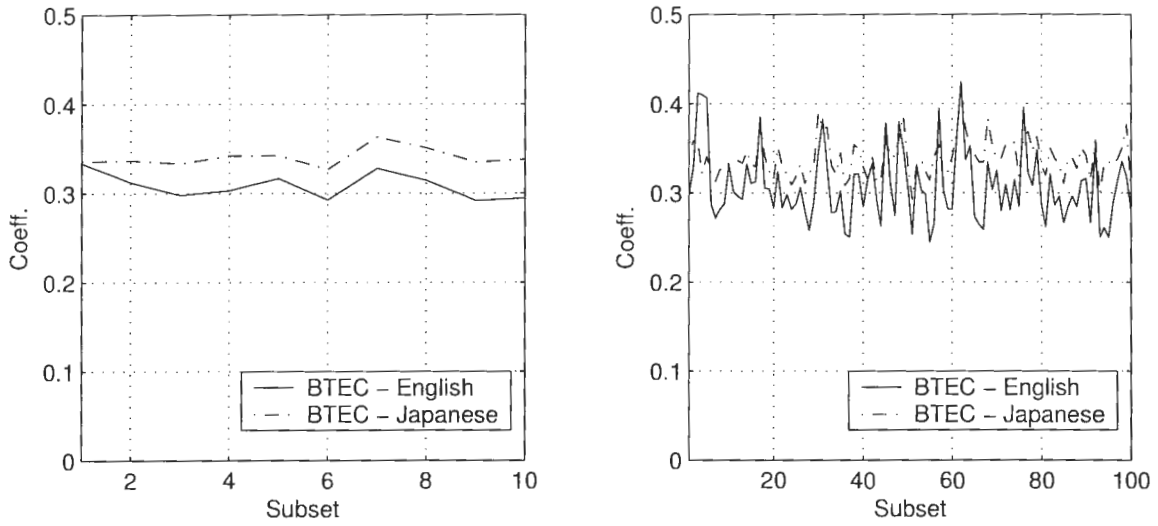

Figure 9: Literality coefficient variation within the BTEC in English and Japanese, respectively for 10 and 100 subsets. As previously the coefficient is computed on 5-gram character models.

Figure 10 shows the correlation between English and Japanese values, and the standard deviation in both languages for an increasing number of subsets.

Subsets	10	50	100	500	1000
Correlation	0.59	0.61	0.70	0.69	0.67
Std dev. Eng	0.014	0.031	0.053	0.063	0.075
Std dev. Jpn	0.008	0.160	0.022	0.031	0.037

Figure 10: Correlation and standard deviations for an increasing number of subsets.

Correlation between English and Japanese values appears to rise and stagnate around a moderately high value of 0.7, the optimum being found at an order of around 100 subsets (each subset therefore containing roughly 1 600 sentences); which tends to

show that indeed, literality in both languages of the same ressource does tend to vary in a similar way, although differences inherent to the languages and their structure rule out the possibility of ever exceeding a certain maximum value in correlation. A more appropriate and accurate subdivision of the corpus remains to be investigated: knowing that the BTEC is a concatenation of sentences taken from various travel phrasebooks, one might assume that there could be differences in style between the various handbooks.

On the other hand, standard deviation provides us with a useful quantification of overall homogeneity, in that it accounts for the average intra-corpus variations of the literality coefficient. As predicted, standard deviation increases as the number of subsets increases and their sizes diminish.

5 Discussion and future work

This work does not state a way in that orality and literality should be defined in an absolute way. It suggests a way to rank corpora in their similarity to each other on a user-definable scale. Ranking is automatic, fast, and does not rely on the counting of any linguistic feature. Classification is only relative to the choice of references, and therefore should be task-oriented. It is clear that this method of evaluating similarity is not limited to the sole application of quantifying literality. A wise choice of references should prove its capacity to rank other criteria correctly. This should be the subject of a future study.

Different corpora should have similar conventions in their levels of transcription and punctuation to be fairly and impartially compared, and to avoid biaised results when dealing with trancripts of oral conversations compared to strict written conventions. An approach at corpus homogeneity in the case of a large multilingual ressource is proposed here, opening the way to a future study on the general quantification of homogeneity whithin a large corpus.

Let us point out once again that it is critical that computational linguistics and corpus-based machine translation have more measures at their disposal for comparing corpora and profiling very large datasets, as we should be able to make a faster and finer link between system performance and corpus features. This is the key to a better understanding of how a corpus-based system behaves, and how to isolate and port its qualities to other systems.

Conclusion

By defining a coefficient of similarity between corpora, and applying it to the differenciating and ranking of literality, we have both tested and confirmed our assumptions on corpora of contrasting sources, having the reputation of being oral or literal. Ranking is robust to N-gram model order variations and is more contrasted for values of N between 3 and 7. We have shown here a general way of relatively classifying and ranking corpus similarity, with the user being free of chosing his own references. This opens

the way to a task-oriented characterisation of corpora, allowing a better understanding and porting of corpus based systems.

Acknowledgements

This research was supported in part by the National Institute of Information and Communications Technology.

References

[Kilgarriff and Rose 1998] Adam Kilgarriff and Tony Rose. 1998.
Measures for corpus similarity and homogeneity.
Proceedings of the 3rd conference on Empirical Methods in Natural Language Processing, Granada, Spain, pp. 46 - 52.

[Kilgarriff 2001] Adam Kilgarriff. 2001.
Comparing Corpora.
International Journal of Corpus Linguistics 6:1, pp. 1-37.

[Biber 1988] Douglas Biber. 1988.
Variation across speech and writing.
Cambridge University Press.

[Biber 1995] Douglas Biber. 1995.
Dimensions in Register Variation.
Cambridge University Press.

[Liebscher 2003] Robert A. Liebscher. 2003.
New corpora, new tests, and new data for frequency-based corpus comparisons.
Center for Research in Language Newsletter, 15:2.

[Teahan and Cleary 1997] William J. Teahan and John G. Cleary. 1997.
Applying compression to natural language processing.
Submitted to ANLP'97.

[Dunning 1994] Ted Dunning. 1994
Statistical identification of language.
CRL New Mexico State University.

[Charniak 1993] Eugene Charniak. 1993
Statistical Language Learning.
MIT Press.

[Nakamura et al., 1996] A. Nakamura and S. Matsunaga and T. Shimizu and M. Tonomura and Y. Sagisaka. 1996
Japanese Speech Databases for Robust Speech Recognition.
Proceedings of the ICSLP'96. Philadelphia, PA, pp.2199–2202, Volume 4

[Matsumoto et al., 2002] Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, Yoshitaka Hirano, Hiroshi Matsuda,Kazuma Takaoka and Masayuki Asahara. 2002
Morphological Analysis System ChaSen version 2.2.9 Manual.
Nara Institute of Science and Technology.

[Sproat and Emerson 2003] Richard Sproat and Thomas Emerson. 2003
The First International Chinese Word Segmentation Bakeoff.
The Second SIGHAN Workshop on Chinese Language Processing, Sapporo, Japan.

A Appendix

This appendix contains abbreviations and typical utterances of the corpora used through this work.

A.1 Summary of English corpora

- SLDB: Spontaneous Speech Database.

  Okay, I go four blocks down Mason Street and then I take a left there, is that right?

- MAD: Machine-translation-Aided bilingual spoken Dialogue corpus .

  Walk two blocks down this street and turn left and you'll see the bank on the right.

- BTEC: Basic Traveller's Expressions Corpus .

  Please send it to this address, if you find my luggage.

- SPAE: The Corpus of Spoken Professional American-English.

  I have carefully read and heard about all of the things that the group has discussed up until now.

- TIME: The TIME Corpus.

  The French, who got no help from the US in developing their force de frappe, were quick to crow that Britain's vaunted ties with the US had brought it nothing but humiliation.

- CALGARY: The Calgary Corpus (book1 & book2 subsets).

  She turned her head to learn if the waggoner were coming.

A.2 Summary of Japanese corpora

- SLDB: Spontaneous Speech Database.

  えーっすぐ分かるんでしょうか、場所は。

- MAD : Machine-translation-Aided bilingual spoken Dialogue corpus .
  すいませんが、写真撮っていただけますか。
- BTEC : Basic Traveller's Expressions Corpus .
  予約をキャンセルしたいのですが。
- NHK : The NHK News Corpus.
  各支店では、行員が新しい仕事の進め方を学ぶ勉強会を開いてきました。
- MAINICHI : The Mainichi Shinbun Corpus.
  ほとんどの企業がその後に五輪競技施設や土木工事を受注していた。
- NIKKEI : The Nikkei Corpus.
  当時、店内には閉店の準備をしていた従業員約二十人がいたが、ほかにけが人はなかった。