Using Word Familiarities and Word Associations to Measure Corpus Representativeness

Reinhard Rapp
Aix-Marseille Université, Laboratoire d'Informatique Fondamentale
163 Avenue de Luminy, 13288 Marseille, France
reinhardrapp@gmx.de

Abstract

The definition of corpus representativeness used here assumes that a representative corpus should reflect as well as possible the average language use a native speaker encounters in everyday life over a longer period of time. As it is not practical to observe people's language input over years, we suggest to utilize two types of experimental data capturing two forms of human intuitions: Word familiarity norms and word association norms. If it is true that human language acquisition is corpus-based, such data should reflect people's perceived language input. Assuming so, we compute a representativeness score for a corpus by extracting word frequency and word association statistics from it and by comparing these statistics to the human data. The higher the similarity, the more representative the corpus should be for the language environments of the test persons. We present results for five different corpora and for truncated versions thereof. The results confirm the expectation that corpus size and corpus balance are crucial aspects for corpus representativeness.

Keywords: corpus representativeness, language intuitions, word familiarities, word associations

1. Introduction

A text corpus can be hoped to be representative of many characteristics of a text, such as of the language of a region, of a particular group of speakers, of a time period, of a subject area, or of a genre. In this paper we propose a method for measuring corpus representativeness. This method could in principle be applicable to any of the above cases. But it requires gold standard data which in practice tends to be only available for general language as typically encountered by native speakers.

The gold standard data we require are such speakers' language intuitions. As our starting point we assume that human language acquisition is essentially corpus-based. By this we mean that when acquiring a language, our brain unconsciously analyzes and stores the statistical properties of the language input we receive. During language production, these properties are reproduced by accessing the stored information. In certain experimental settings it is possible to focus on particular aspects of language production. This way some of the stored information can be accessed specifically.

Previous work provides evidence that this is possible at least for the following two types of information: Word frequency (Rapp, 2005) and word association (Wettler et al., 2005).[1] Concerning word frequencies, although people are generally not aware that their brain stores information on the occurrence frequencies of the words they perceive, it can be shown in two ways that this is nevertheless the case: The reaction times for the word recognition task[2] can be measured, or test persons can be asked to rate word familiarities. In both cases there is a high correspondence with word frequency data as taken from balanced corpora such as the British National Corpus (Rapp, 2005; Brysbaert & New, 2009).

Concerning word associations, it has been found that these can be derived from text corpora. The results show a high level of similarity to human word associations, and the procedure used is in agreement with psychological learning theory (Wettler et al., 2005).

Whereas previous work had the focus on optimizing the algorithms with the aim of maximizing the similarity between the corpus-derived and the human associations, in the current paper we change the perspective: We take the algorithm as fixed and vary the corpora. We then say that the better the corpus derived associations match the human associations, the higher should be the representativeness of the respective corpus. Hereby we use as our gold standard compilations of human intuitions as collected from test persons in large scale experiments.

In the following sections, after describing the nature of the human intuitions, we realize this approach for the above two types of information, namely word frequency and word association. In Rapp (in print) and Rapp (2014) we have dealt with the two aspects before, but here we for the first time compare the two approaches.

2. Human language intuitions

2.1 Word familiarities

The first type of human intuitions which we would like to consider are word familiarities as obtained from test persons. Such data has been experimentally collected by psychologists from native speakers, as exemplified in the *MRC Psychological Database* (Coltheart, 1981) where familiarities for thousands of test words are listed.

[1] Another such property is word relatedness in the sense of Harris' distributional hypothesis (Pantel & Lin, 2002). We are currently conducting work analogous to what is described in this paper for word-relatedness, with roughly similar outcome.

[2] Asking test persons questions such as: Is *table* an English word? Or: Is *elbat* an English word?

To collect the familiarity judgments, test persons were asked to rank the subjective familiarities of words on a scale between 1 and 7. Hereby, 1 is to be assigned to an unknown word, and 7 to a very familiar word from everyday language. The familiarity judgements resulting from such experiments have been compiled in large tables, the so called *familiarity norms*. For our experiments, we decided to use the familiarity norms included in the MRC Psycholinguistic Database which is actually a conglomerate of three familiarity norms comprising altogether 4920 words.

In previous work (for an overview and references see Rapp, 2005) it has been shown that there is some correspondence between the human familiarity judgments and the corpus frequencies of words in text corpora. For illustration, Table 1 shows the top six most familiar words in the MRC database together with their frequencies in the Brown corpus and compares them to the least familiar words. As can be seen, the familiar words have consistently much higher corpus frequencies on average.

Rapp (2005) reports a high correlation (according to Pearson) between the subjects' familiarity judgements and the logarithm of the observed corpus frequencies. As an explanation it is hypothesized that human familiarity ratings are based on the word frequencies as observed by the test persons in the everyday language they perceive.

However, if the familiarity norms reflect word frequencies in perceived language, then it should be possible to use them as a standard for measuring the frequency aspect of corpus representativeness. A corpus whose word frequencies are highly correlated to the familiarity norms is likely to be a good surrogate for everyday language, although word frequency of course reflects only one of many properties of a corpus, see the reflections on this in Section 7. Nevertheless, for a corpus to be representative, it is a necessary (though not sufficient) condition that its word frequencies are similar to those in everyday language.

WORD	FAM.	FREQ.
BREAKFAST	6.6	53
AFTERNOON	6.5	106
CLOTHES	6.5	89
BEDROOM	6.5	52
DAD	6.5	15
GIRL	6.5	220

WORD	FAM.	FREQ.
LOQUACITY	1.4	1
MIEN	1.4	1
YUCCA	1.4	1
BURGHER	1.3	1
PAEAN	1.3	2
OBELISK	1.3	6

Table 1: The six words with the highest and the lowest familiarities in the MRC Psycholinguistic Database together with their frequency counts in the Brown Corpus (words with a corpus frequency of zero are not included).

2.2 Word associations

The other type of human intuitions which we would like to consider are word associations as obtained from test persons. Such data has been collected from native speakers in large scale experiments, as exemplified in the *Edinburgh Associative Thesaurus* (*EAT*; Kiss et al., 1973) where English word associations for thousands of stimulus words are listed.

It is relatively straightforward to conduct such experiments: Typically, the subjects are given questionnaires with lists of stimulus words, and are asked to write down for each stimulus word the spontaneous association which first comes to mind. This leads to collections of associations, the so-called association norms, as exemplified in Table 2.

ABOVE	CONSTELLATİON	FEMİNİNE
below (59)	stars (39)	masculine (26)
high (4)	star (33)	girl (14)
over (4)	sky (5)	woman (8)
sky (4)	andromeda (2)	female (6)
all (3)	aquarius (2)	sex (3)
up (3)	plough (2)	beauty (2)
me (2)	aircraft (1)	bird (2)
under (2)	cancer (1)	girls (2)
us (2)	clear (1)	nice (2)
average (1)	dance (1)	pretty (2)

Table 2: Top ten associations to three stimulus words as taken from the EAT. The numbers of subjects responding with the respective word are given in brackets.

Association theory, which can be traced back to Aristotle in ancient Greece, has often stated that our associations are governed by our experiences. Now the question arose whether for perceived words the same principles might apply, and with the advent of corpus linguistics it was possible to verify this experimentally by looking at the distribution of words in texts. Among the first to do so were Schvaneveldt et al. (1989), Wettler & Rapp (1989), and Church & Hanks (1990).

Their underlying assumption was that strongly associated words should often occur in close proximity in text corpora. This is actually confirmed by corpus evidence: Figure 1 assigns to each stimulus word position 0, and displays the occurrence frequencies of its *primary associative response* (most frequent response as produced by the test persons) at relative distances between -50 and +50 words. However, to give a general picture and to abstract away from idiosyncrasies, the figure is not based on a single stimulus/response pair, but instead represents the average of 100 English stimulus/response pairs as published by Jenkins (1970). The effect is in line with expectations: The closer we get to the stimulus word, the higher the chances that the primary associative response occurs. Only the distances plus and minus one and plus and minus three are exceptions, but this is an artefact because content words are typically separated by function words which carry not much content and are therefore of little interest here.

Word associations derived from corpora are the basis underlying the second part of this work. Wettler et al. (2005) linked these to psychological learning theory, thereby providing strong evidence that human association learning is in essence corpus based. A later paper by Turney & Pantel (2010) nicely works out the relationship between frequency and meaning. In short, there is some evidence that the framework assumed here is sound.

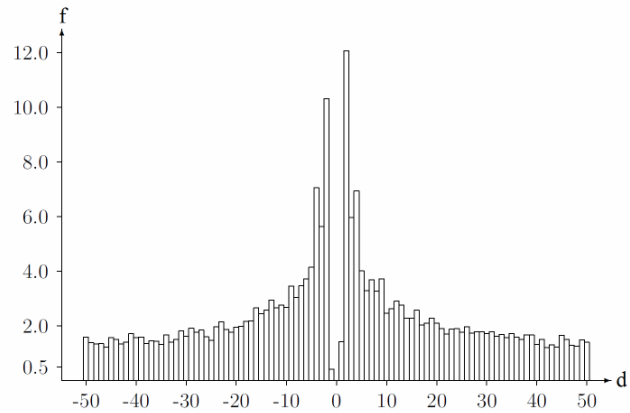

Fig. 1: Occurrence frequency *f* of a primary response at distance *d* from a stimulus word, averaged over 100 stimulus/response pairs. At large distances from the stimulus word the average occurrence frequency of a primary response is 0.49 (Rapp, 1996).

3. Corpora

Our corpus representativeness measure is to be applied to a number of well known corpora. These are:

1. Brown Corpus (balanced corpus of 1 million words; Francis & Kuçera, 1989)
2. British National Corpus (BNC; balanced corpus of 100 million words; Burnard & Aston, 1998)
3. English Wikipedia (300 million words of encyclopaedic texts)[3]
4. ukWaC (British English web corpus of 2 billion words)[4]
5. English Gigaword Corpus 4th edition (4 billion words of newswire text)[5]

Both the MRC database and the EAT use uppercase characters only as at the time of their construction a distinction between uppercase and lowercase characters was not yet standard in computing. As this is only a minor shortcoming, we decided not to try to make up for it. We only converted both resources to lowercase to improve readability. For reasons of consistency, we also converted all corpora to lowercase.

For the experiments described in the next sections we needed to cut off our corpora in order to provide results for subcorpora of particular sizes. Because corpus size was measured as the number of running words, and as these numbers depend somewhat on how word segmentation is conducted, let us briefly describe the procedure we used: In order to keep our algorithm as language independent as possible, we simply consider any uninterrupted sequences of alpha characters as words, but also any sequences of non-alpha characters except white space (blanks, tabulator, new line). That is, white space and transitions between the two types of characters (alpha and non-alpha) are considered as word separators.

[3] We use the English part of the Wikipedia XML Corpus (Denoyer & Gallinary, 2006). Although this is smaller than current versions, it has the advantage that it is an offline copy so that our results can be replicated.

[4] http://wacky.sslmit.unibo.it/doku.php?id=corpora

[5] http://catalog.ldc.upenn.edu/LDC2009T13

4. Procedure

4.1 Corpus statistics concerning familiarities

In the case of word familiarities the statistics extracted from the corpora are very simple, namely the log frequencies of the words. The MRC database contains familiarities for 4920 words. As just two of them are multiword units, we considered this an inconsistency and removed them, so that 4918 words remained. Another reason was that multiword units might possibly require different treatment than single words. For example, it is not clear how the familiarities of the components combine to the familiarity of the entire multiword unit.

Word	Word frequency in the BNC	Word familiarity in the MRC database
a	2247100	632
abandon	1316	510
abandonment	500	359
abasement	20	226
abatement	137	294
abbess	57	187
abdication	124	284
abdomen	303	426
abduction	230	413
aberration	149	208
abhor	43	360
ability	9190	563
able	30634	575
abode	167	334
abominable	88	358
aboriginal	222	295
about	198665	593
above	25935	607
abridgement	34	303
abrupt	499	486

Table 3: BNC frequencies and MRC familiarities for the (alphabetically) top 20 words covered in the familiarity norms of the MRC database.

The two types of data, namely the familiarity norms from the MRC database and the word frequencies as extracted from one of the corpora, were merged as exemplified in Table 3 for the case of the BNC. Note that although the test subjects' familiarity judgements were originally on a scale between 1 (not familiar) and 7 (highly familiar), to avoid decimal numbers when averaging results, all ratings were multiplied by 100 which is reflected in the table.

Computing a corpus representativeness score now simply involves taking the logarithm (base 10) of the frequencies in column 2, and then computing Pearson's correlation coefficient between the resulting vector and column 3. However, as especially for small corpora many of the word frequencies can be zero, and as the logarithm of zero is not defined, we applied the usual heuristic of adding one to each number before taking the logarithm.

4.2 Corpus statistics concerning associations

As discussed previously, in this section we assume that there is a relationship between word associations as collected from human subjects and word co-occurrences as observed in a corpus, and our hypothesis is that the strength of this relationship can be used as a measure of corpus representativeness. A corpus leading to simulated associations very similar to the ones collected from humans is likely to be a good surrogate for everyday language, although word associations constitute only one of many properties of a corpus. Nevertheless, for a corpus to be representative, it is a necessary (though not sufficient) condition that the word associations derived from it are similar to those collected from humans.

As our source of human data we use the EAT (Kiss et al. 1973) which is the largest classical collection of its kind. The EAT comprises the associative responses as requested from around 100 British students for each of altogether 8400 stimulus words. As exemplified in Table 2, some of the responses to a particular stimulus word are given by many students, whereas others are given by only one or two. What is the reason for this? According to the theory, the associative response of a test person should reflect this persons's language environment, i.e. the text and speech the person previously encountered and whose statistical properties were stored in long term memory. As apparently there is some variation in each person's language history, some variation in the associative responses can be expected. This is a natural and desired effect.

However, there is also an undesired component in this variation: That test persons perceive only a single input word at a time is an idealization. In reality, they still have the previous input words in short term memory, and also the (e.g. classroom) environment provides lots of additional (e.g. visual) stimuli. So what the subjects actually come up with is based on a mix of all these stimuli plus of what they have in short (and possibly medium) term memory. For example, the responses on the stimulus word *artist* could be influenced by some artwork present in classroom, or of a museum visit the day before. The responses to *transport* could be influenced by the means of transport the students used for arriving in class, such as car, bicycle, train, or bus. Or the response to *cinema* could depend on a recently watched movie.

As corpus representativeness should focus on long term averages and not on short term effects, such influences are undesired and should be avoided. We do so by assuming: Responses confirmed by many students are more likely to reflect long term averages (especially if the group of test persons is heterogeneous). And responses provided by only one or two test persons are more likely to be noise.

As the EAT is rather large, we can afford to stay on the high quality side of this scale. For this reason, we decided to use only the primary associative response for each stimulus word, and to discard all other responses. Alternatively, we could have decided to take the top two or top five responses into account. However, preliminary experiments showed that this would "water down" our results. That is, the basic effects would remain the same, but in a somewhat less salient fashion. Compare Rapp (2013) where such effects are quantified for the related task of multiword association.

Like the MRC familiarity norms, the EAT also contains some multiword units. A problem is that when computing associations for single word stimuli, it is not clear whether or not matching components of a multiword unit should be considered. For example, given *New Year*, should the occurrences of *Year* within this multiword unit nevertheless be used for computing the associations to *Year*? As for computing corpus representativeness it seemed not important to elaborate on this less relevant problem, we simply decided to remove all items from the EAT where either the stimulus word or the primary associative response happened to be a multiword unit.

Finally, as high frequency function words were considered of little relevance to our analysis and in order to keep our algorithm efficient, we decided not to take into account some of the most frequent words. Based on the word frequencies in the British National Corpus we compiled a list of 48 words with frequencies 250,000 or higher. If either the stimulus word or its primary response occurred in this list, we removed the respective item from the EAT.

In summary, of the 8400 items in the EAT we removed those involving multiword units and high frequency function words, thereby obtaining a list of 7731 remaining items.

Whereas some previous studies involving the computation of word associations used lemmatization, we decided not to do so here. Firstly, in view of future work, we wanted to keep the basic algorithm for computing word associations as language independent as possible. Secondly, lemmatizing the EAT is problematic as it does not provide context for its words. Thirdly, lemmatization has benefits mainly at the evaluation stage, but this is not important in our setting. For example, given the stimulus word *table* and the primary associative response *chair*, if the simulation produced the plural form *chairs* this would count as incorrect without lemmatization. But as this is a relatively infrequent phenomenon, and as it applies in a similar way across corpora, it is not very relevant for corpus comparisons where the focus is not on polishing individual results.

For extracting word associations from our corpora we used the following procedure: For all words occurring as stimuli in our EAT derived gold standard we computed the co-occurrence vectors. That is, each vector contains the number of co-occurrences of the stimulus word with

all other co-occurring words. It counts as a co-occurrence if two words appear together within a distance of at most ten words, i.e. a text window of plus and minus 10 words around the stimulus word is considered. Hereby the exact distance within the window is not taken into account.

Having completed the co-occurrence counting, in the next step an association measure was applied to the co-occurrence vectors. This is meant to account for the differences in absolute word frequencies. As our association measure of choice we used the log-likelihood ratio (Dunning, 1993) which is very well established for such purposes. It compares the observed co-occurrence counts with the co-occurrence counts expected from chance, thus strengthening significant word pairs and weakening incidental word pairs. The resulting vectors we call association vectors. Given these vectors, the strongest association to a given stimulus word can be determined by simply looking for the highest value within the respective association vector. The corresponding word is considered to be the associative response predicted by the system. For the same stimulus words used in Table 2, Table 4 shows some sample associations as computed using the British National Corpus.

Let us briefly discuss the above choice of window size (±10 words). Some previous studies have used smaller window sizes such as e.g. ±2 words in Wettler et al. (2005). However, other than in the present work this study had eliminated function words from the corpus, so that the effective window size might have been around ±4 words. Also, it should be noted that the best window size depends on the size of the corpus: For small corpora the problem of data sparseness can be somewhat reduced by considering a larger window, whereas for larger corpora this is not necessary and a smaller window might possibly lead to a higher accuracy of the predictions (compare Figure 1).

ABOVE	CONSTELLATİON	FEMİNİNE
below (59)	stars (39)	masculine (26)
level	star (33)	women (2)
average (1)	southern	gender
high (4)	triangle	woman (8)
feet	bright	female (6)
water	planet (1)	men
head	rather	male (1)
see	south	more
ground	find	hair
left	map	soft

Table 4: Top ten corpus-derived associations for three stimulus words. The numbers of subjects from the EAT responding with the respective word (if larger than zero) are given in brackets.

Concerning evaluation, in principle the idea is to find matches between the human and the corpus-based associations. One possibility is to simply count the number of cases where the primary associative response matches the strongest corpus-based association. However, when it comes to very small corpus sizes of e.g. just 1000 words (see Section 5), the problem of data sparseness becomes so severe that a more tolerant evaluation method leads to more robust results less susceptible to statistical variation. This is why for measuring accuracy we count the number of cases where the primary associative responses is listed within the top ten corpus-based associations, rather than insisting on a match with the strongest association. This simple modification leads to improvements in reliability when measuring very low accuracies.

As some readers may expect evaluations based on recall, precision, and/or f-measure, let us explain why we believe that these are not very appropriate here. In principle, it would be possible to e.g. look at the top ten human associations for a given stimulus word, and then find out how many of these occur in the top ten corpus-based associations. From the results recall, precision and f-measure could be computed. However, the problem is the following: These measures were developed in Information Retrieval under the assumptions that within the documents in a database two categories can be distinguished: Those relevant to a query and all others which are assumed to be irrelevant. However, what we have in the case of word associations is that the degree of relevance (here: association strength) is very important. For example, given the stimulus word *black*, 57 of 99 subjects answered with *white*, but each of the following 30 responses is given by at most three subjects. The problem is that an evaluation based on recall and precision would give such spurious responses the same weight that it gives the top response, which is clearly inappropriate. In short: We have chosen our straightforward evaluation methodology not because it is simpler, but because it is considerably better for this particular purpose.

5. Results

Concerning the representativeness of our five corpora, we tried to come up with some hypotheses before we started to compute the results. These were our predictions:

1. Representativeness should increase with corpus size.
2. The Brown corpus and the BNC should be more representative than unbalanced corpora of the same size.
3. The Brown corpus (1 million words) should be more representative than the first million words of the British National Corpus as the latter is balanced only over its full size (100 million words), but not over its first million words.
4. For same sizes, we would expect ukWaC to be more representative than Wikipedia as we think that corpus heterogeneity is a plus for representativeness. ukWaC is obviously more heterogeneous as, for example, it is multi genre multi topic whereas Wikipedia is single genre multi topic.
5. The Gigaword Corpus should be the least representative for identical sizes. Although, like Wikipedia, it is also single genre multi topic, the distribution of topics is not as wide because in newsticker texts there are strong foci e.g. on politics and sports.

The actual results are given separately for the two approaches in the following two subsections.

5.1 Results based on word familiarities

These results are given in Table 5. There we find for each of the five corpora its size and the computed Pearson correlation coefficients between the MRC words' familiarities and their corpus frequencies. For better comparison with the results presented in the next section (which are percentages) we multiply these correlations by 100 and take this product as the *representativeness of a corpus*. The range of values can thus be between 0 and 100, whereby 0 denotes a complete lack of representativeness, and 100 denotes perfect representativeness. The representativeness scores are also computed for partial corpora, whereby all parts have in common that they start with the beginning of the respective corpus.

Corpus size (words)	Brown	BNC	Wiki-pedia	uk-WaC	Giga-word
100	7.63	9.02	8.15	9.44	3.29
1000	15.94	20.09	15.75	18.76	14.88
10000	29.66	32.80	27.68	32.71	29.86
100000	49.48	48.67	46.70	50.57	44.22
1 million	67.04	68.50	59.89	64.40	55.00
10 million	–	73.61	65.53	71.37	62.20
100 million	–	75.13	66.66	73.31	66.00
1 billion	–	–	–	73.68	69.10
full corpus	68.89	75.56	67.04	73.72	73.35
full corpus mill. words	1.18	117	313	2345	4371

Table 5: Familiarity-based corpus representativeness for full and partial corpora. Representativeness is measured as the correlation between the corpus frequencies of words and their familiarities. For easier comparison with Table 6 all correlations were multiplied by 100.

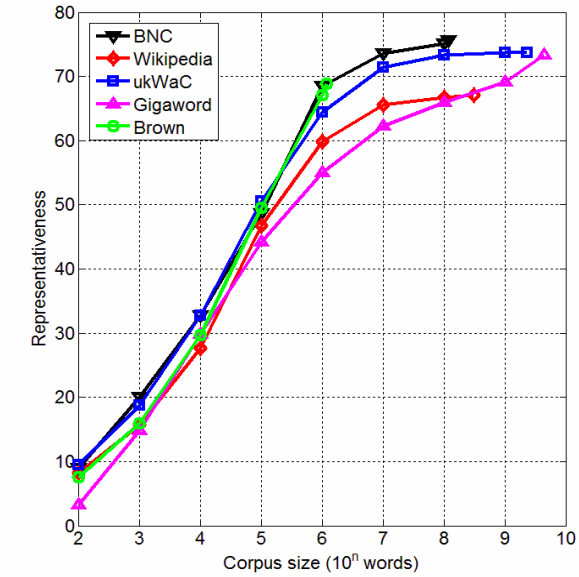

Fig. 2: Pearson's correlation coefficient (×100) between corpus frequencies and word familiarities depending on corpus size for five corpora.

We can see in Table 5 that, as expected, the representativeness is almost zero if only the first 100 words of a corpus are taken into account, and gradually increases to at least 67 for the full corpora. This increase can be better seen in Figure 2 which is a graphical representation of Table 5. The horizontal axis has a logarithmic scale, but still the curves flatten with increasing corpus size, especially above 1 million words. Note that the curve for the Brown corpus is not very well visible as it ends at 1 million words and thus has a range where all curves are overlapping.

5.2 Results based on word associations

These results are given in Table 6, with a graphical representation provided in Fig. 3. There we find for each of the five corpora its size and the percentage of primary associative responses which ranked among the top ten in the corpus-based associations. These percentages we take as the association-based representativeness of the respective corpus. The range of values can be between 0 and 100, whereby 0 denotes a complete lack of representativeness, and 100 denotes perfect representativeness. The values are also provided for partial corpora, whereby the parts always start at the beginning of the respective corpus.

Corpus size (words)	Brown	BNC	Wiki-pedia	uk-WaC	Giga-word
100	0.31	0.22	0.30	0.47	0.13
1000	0.49	0.67	0.72	0.61	0.28
10000	0.60	0.91	0.71	0.69	0.83
100000	2.95	3.09	3.12	2.90	1.91
1 million	13.22	13.85	11.24	11.87	5.91
10 million	–	35.26	26.83	30.02	14.20
100 million	–	52.67	42.89	48.40	25.13
1 billion	–	–	–	55.92	35.57
full corpus	14.62	53,63	49.89	57.07	44.43
full corpus mill. words	1.18	117	313	2345	4371

Table 6: Association-based corpus representativeness scores for full and partial corpora.

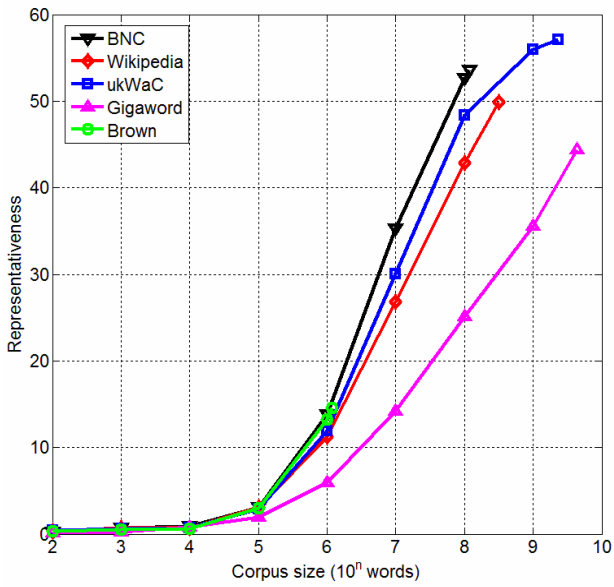

Fig. 3: Association-based corpus representativeness scores depending on corpus size.

6. Discussion

If we compare the curves in Figures 2 and 3, it is apparent that the shapes are rather different. This can be explained by obvious differences in the two methods: The familiarity-based approach uses statistics of order zero (word frequencies), whereas the association-based approach uses first order statistics (word co-occurrences). Although for both methods a flattening of the curves can be expected for large corpora for the reason that there is an upper limit of corpus representativeness (100) leading to saturation, apparently for the first order statistics even larger corpora would be needed to make this happen.

Concerning very small partial corpora, for the familiarity based approach the curves quickly rise, whereas for the association-based approach the increases in accuracy are small at the beginning. This is also to be expected because in a partial corpus of e.g. 1000 words only few of the 7731 EAT stimulus words can occur, and even fewer of the expected co-occurrences.

So these discrepancies between the two approaches are not a surprise. They are roughly analogous to what could be expected when considering the frequencies of n-grams of different lengths. Of more interest is a comparison of the results for the different corpora, i.e. their relative performance. In this respect we can see quite some similarities, which is desirable as the claim is that both methods are supposed to measure aspects of the same thing, namely corpus representativeness.

In particular, if we consider the hypotheses stated in Section 5, the findings are as follows:

Hypothesis 1, namely that the representativeness of all corpora steadily increases with corpus size is clearly confirmed by both approaches.

Hypothesis 2, saying that the balanced corpora, namely the Brown corpus and the BNC, should be more representative for their sizes than non-balanced corpora, is also confirmed by both approaches. At 1 million words, these two are the top performers. At 100 million words, the BNC performs best. Note, however, that the smaller the corpus sizes, the less predictable the results as the sampling errors increase. For this reason it probably does not make much sense to compare the representativeness scores among the smaller partial corpora (e.g. below 100,000 words). We only presented them here to be able to verify hypothesis 1.

Hypothesis 3 (Brown better than BNC for 1 million words) could not be confirmed by any of the two approaches, although the results are fairly close in both cases. Our explanation for this is that the BNC better matches the language environment of the EAT students (experiments were conducted in Edinburgh) as it represents British English whereas the Brown corpus represents American English. Apparently this effect is stronger than the balancing effect we had in mind.

Possibly the time periods when the text samples were produced might also play a role. The EAT experiments were conducted between June 1968 and May 1971. The BNC's text samples mainly date from 1975 to 1994 (only some imaginative texts date back earlier: 1960 to 1974). The materials in the Brown corpus were all published in the United States in 1961. This means: The EAT students' language environment was pre 1968 to pre 1971. The BNC authors' was mainly pre 1975 to pre 1994. And the Brown authors' pre 1961.

Hypothesis 4, namely that ukWaC is better than Wikipedia, is confirmed for all corpus sizes above 100,000 words. As noted before, for smaller corpus sizes sampling errors are likely to be significant.

Hypothesis 5, saying that the Gigaword corpus should be the least representative, is clearly confirmed for almost all corpus sizes.

Overall four of the five hypotheses were confirmed by both methods, and the other hypothesis was rejected by both for a good reason that we had overlooked. This provides some evidence that the computed scores are actually related to what might sensibly be considered as the representativeness of a corpus.

In particular, it is worth noting that our measure confirms the intuition that it makes sense to balance a corpus and that corpus heterogeneity is a plus: The balanced BNC performs best at 100 million words, and the very heterogeneous ukWaC is another top performer.

7. Summary and outlook

In this work we defined the term *corpus representativeness* as the ability of a corpus to represent the average language use a native speaker encounters in everyday life. As we cannot easily observe test persons over years, our suggestion was to utilize human intuitions on word familiarities and on word associations.

Previous work had provided evidence that human word familiarities are based on word frequencies in perceived language, and that human word associations are based on the co-occurrences of words in perceived language. Although this may still be controversial, in the current work we took these findings for granted but turned round the perspective. We said that a corpus is representative for the language environment of a group of persons if the word familiarities and word associations derived from it resemble these persons' intuitions.

For full and partial versions of five well known English corpora we computed the word familiarities and word associations for test sets of several thousand words. We then, for each corpus, compared the resulting familiarities and associations to the human data, and computed similarity scores which we took as measures of corpus representativeness.

Note that our measures only aim for representativeness concerning an average person's language environment (rather than e.g. the sum of the language environments of all speakers of a language). Therefore it is by design that it does not take into account a corpus' comprehensiveness beyond the vocabulary and associations an average native speaker knows. Therefore it can be justified that the very large Gigaword corpus, despite its rather comprehensive vocabulary, did not perform well.

But considerably more severe appears the following shortcoming: Our measures are limited in so far as they only consider two particular aspects of corpus represent-

ativeness, namely word familiarity and word association. They do not explicitly consider higher level features e.g. concerning syntax, semantics, pragmatics, or style.

Let us therefore, as a next step, propose an extension to be dealt with in future work: Whereas word familiarities and word associations are based on statistics of orders zero and one, we suggest to extend the method to statistics of order two. Second order statistics concern word relatedness, thereby – in the spirit of Harris' distributional hypothesis – identifying words with common context. Here human data is also available: A prototypical example is the data from the well known TOEFL synonym test (Landauer & Dumais, 1997). But synonym dictionaries could also be considered as human data, and in particular WordNet whose variants are available for many languages. That is, the quality of corpus-derived WordNet synsets could be taken as a measure of corpus representativeness.

After compiling a number of scores representing statistics of order zero, one and two, these scores might finally be combined into an overall score e.g. by computing their geometric mean. This would be in analogy to the BLEU score (Papineni et al., 2002) used in machine translation evaluation where matches (between machine translation and human reference translation) of various n-gram lengths are separately scored and then combined.

Acknowledgment

This research was supported by a Marie Curie Intra European Fellowship within the 7th European Community Framework Programme.

References

Biber, D. (1993): Representativeness in Corpus Design. *Literary and Linguistic Computing*, Vol. 8, Nov. 4, 243–257.

Brisbaert, M.; New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods 41 (4),* 977–990.

Burnard, L.; Aston, G. (1998): *The BNC Handbook: Exploring the British National Corpus with Sara.* Edinburgh: University Press.

Church, K.W.; Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics* 16 (1), 22–29.

Coltheart, M. (1981): The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, 33A, 497–505.

Denoyer, L.; Gallinari, P. (2006): The Wikipedia XML Corpus. *SIGIR Forum*, 40(1), 64–69.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistic*s, 19 (1), 61–74.

Francis, W.N.; Kuçera, H. (1989): *Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English, for Use with Digital Computers.* Providence, R.I.: Brown University, Department of Linguistics.

Jenkins, J.J. (1970). The 1952 Minnesota word association norms. In: L. Postman; G. Keppel (eds.): *Norms of Word Association*. New York: Academic Press, 1-38.

Kiss, G.R., Armstrong, C., Milroy, R., and Piper, J. (1973). An associative thesaurus of English and its computer analysis. In Aitken, A.J., Bailey, R.W. and Hamilton-Smith, N. (Eds.): *The Computer and Literary Studies*. Edinburgh: University Press, 153-165.

Landauer, T.K.; Dumais, S.T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* 104 (2), 211-240.

McEnery, T.; Wilson, A. (1996): *Corpus Linguistics*. Edinburgh University Press.

Pantel, P.; Lin, D. (2002). Discovering Word Senses from Text. In: *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Edmonton, 613–619.

Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the ACL*, Philadelphia, 311-318.

Rapp, R. (2005): On the relationship between word frequency and word familiarity. In: B. Fisseni; H.-C. Schmitz; B. Schröder; P. Wagner (eds.): *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen. Beiträge zur GLDV-Tagung 2005 in Bonn.* Frankfurt: Peter Lang. 249–263.

Rapp, R. (2013): From stimulus to associations and back. *Proceedings of the 10th Workshop on Natural Language Processing and Cognitive Science*, Marseille, France.

Rapp, R. (in print). Using word familiarities to measure corpus representativeness. *Akten des 48. Linguistischen Kolloquiums*, Alcala, Spanien, 2013.

Saldanha, G. (2009): Principles of corpus linguistics and their application to translation studies research. *Tradumática* 7: 1–7.

Schvaneveldt, R. W., Durso, F. T., & Dearholt, D. W. (1989). Network structures in proximity data. In G. Bower (ed.): *The Psychology of Learning and Motivation: Advances in Research and Theory,* Vol. 24. New York: Academic Press, 249–284.

Turney, P.T.; Pantel, P. (2010). From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research* 37, 141-188.

Wettler, M., Rapp, R. (1989). A connectionist system to simulate lexical decisions in information retrieval. In: R. Pfeifer, Z. Schreter, F. Fogelman, L. Steels (eds.): *Connectionism in Perspective*. Amsterdam: Elsevier, 463–469.

Wettler, M.; Rapp, R.; Sedlmeier, P. (2005). Free word associations correspond to contiguities between words in texts. *Journal of Quantitative Linguistics* 12(2), 111–122.