# Using Collections of Human Language Intuitions to Measure Corpus Representativeness

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

## Abstract

In corpus linguistics there have been numerous attempts to compile balanced corpora, resulting in text collections such as the Brown Corpus or the British National Corpus. These corpora are meant to reflect the average language use a native speaker typically encounters. But is it possible to measure in how far these efforts were successful? Assuming that humans' language intuitions are based on our brain's capability to statistically analyze perceived language and to memorize these statistics, we suggest a method for measuring corpus representativeness which compares corpus statistics to three types of human language intuitions as collected from test persons: Word familiarity, word association, and word relatedness. We compute a representativeness score for a corpus by extracting word frequency, word co-occurrence, and contextual 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. Our findings confirm the expectation that corpus size and corpus balancing matter.

## 1 Introduction

Balanced corpora, i.e. corpora consisting of a carefully sampled mix of texts, have often been considered important for providing a standard of average language use. Well known examples of such corpora include the *Brown Corpus* (Francis & Kuçera, 1989) and the *British National Corpus* (Burnard & Aston, 1998). But to obtain a balance many decisions concerning the corpus design have to be made. Biber (1993) mentions, among other things, that it has to be decided for what target population a corpus is meant to be representative, that estimates concerning the quantities of various text types are required, and that decisions with regard to the number of individual text samples and their sizes have to be made.

However, there is no easy and well established way to verify the success of these measures. Current suggestions include, for example, to consider a corpus as representative if it is not dominated by sublanguage (Temnikova et al., 2014), or to more or less give up on the concept of representativeness and to concentrate on considering the suitability of a corpus for particular tasks. Saldanha (2009) comes to the conclusion that "The problem with making representativeness the defining characteristic of a corpus is that it is very difficult to evaluate."

Our goal here is to make an attempt to measure corpus representativeness in a standardized way, thereby avoiding to observe test persons' average language input as this would not be very practical. Our starting point is that a representative corpus should reflect as well as possible average language use as encountered by native speakers. We also assume that human language acquisition is essentially corpus-based (Rapp, 2011). This implies the following: The human brain analyzes particular statistical properties of perceived language and memorizes them. During language production these properties





*Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2117–2128, Dublin, Ireland, August 23-29 2014.

are reproduced. It has been shown that in certain test situations it is possible to isolate intuitions related to some specific statistics. These include the following three which we will utilize for measuring corpus representativeness: Word frequency, word co-occurrence, and common context of words. In terms of human language intuitions, these three statistical properties relate to word familiarities, word associations, and word relatedness.

What we suggest is to extract data relating to these three types of statistical properties from a corpus and to compare it to the respective experimental data as obtained from test persons. The higher the average agreement, the more representative the corpus should be for the language environment of the test persons.[1]

Related work has been conducted by Brisbaert & New (2009), which is mentioned in section 2.1, and in our own previous studies (Rapp, 2014a and Rapp, 2014c), of which the current work is an extension. A nice summary of how to measure corpus representativeness through psycholinguistic measures is provided in a presentation by Francom & Ussishkin (2011). Gries (2010), though in a slightly different context, emphasizes the need of external validation: "For corpus linguists, that means that our measures must be validated against corpus-external evidence because, strictly speaking, as long as we corpus linguists do not show that our dispersions and adjusted frequencies correspond to something outside of our corpora, we have failed to provide the most elementary aspect of a new measure – its validation."

The remainder of this paper is structured as follows: We first describe the experimental data used, i.e. the familiarity norms, the association norms, and the synonym data (describing word relatedness). Next we present the algorithms used to extract the corresponding statistics from the corpora. By comparing the human and the corpus-derived data, we introduce three quantitative measures of corpus representativeness, which we subsequently combine. The paper concludes with a discussion and an outlook on future work.

## 2 Human language intuitions

### 2.1 Word familiarities

Psychologists have collected word familiarity ratings from test persons. For this purpose, the subjects were asked to come up with subjective familiarities for given words. Usually a scale between 1 and 7 was used, whereby 1 means unfamiliar and 7 means very familiar. The outcome of such experiments are the so-called familiarity norms, i.e. large tables listing the subjects' familiarity ratings. In the current work we used the familiarity data for 4920 words from an online version[2] of the *MRC Psycholinguistic Database* (Coltheart, 1981).

In previous studies (e.g. Rapp, 2005) it has been shown that there is a strong correlation between the human familiarity judgments and the log occurrence frequencies of the words in corpora. For illustration, Table 1 shows the top five most familiar words in the MRC database together with their frequencies in the Brown corpus and compares them to some of the least familiar words. As can be seen, the familiar words have consistently much higher corpus frequencies. To explain this finding, Rapp (2005) hypothesized that human familiarity ratings are based on the word frequencies as observed by the test persons in the language they perceive in everyday life.

However, if we assume that 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 more likely to be a good surrogate for everyday language, although word frequency of course reflects only one of many properties of a corpus. 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.

[1] Let us mention that there is some analogy to automatic MT evaluation, namely when computing the BLEU score: There a machine translation is compared to a human translation (which is based on human intuitions) by identifying matches between n-grams of various lengths. Then a combined score is computed from the results obtained for each n-gram length.

[2] http://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm

We should mention that instead of using word familiarity data it is also possible to use reaction times as obtained in the word recognition task.[3] Brisbaert & New (2009) did so and related the reaction times to the corpus frequencies of the words for the purpose of measuring corpus representativeness. In essence, although they tested on other corpora, their findings seem to be similar to what we report here based on word familiarities.

| FAMILIAR WORDS | | | UNFAMILIAR WORDS | | |
|---|---|---|---|---|---|
| WORD | FAMILIARITY | BROWN FREQUENCY | WORD | FAMILIARITY | BROWN FREQUENCY |
| BREAKFAST | 6.6 | 53 | LOQUACITY | 1.4 | 1 |
| AFTERNOON | 6.5 | 106 | MIEN | 1.4 | 1 |
| CLOTHES | 6.5 | 89 | YUCCA | 1.4 | 1 |
| BEDROOM | 6.5 | 52 | BURGHER | 1.3 | 1 |
| DAD | 6.5 | 15 | PAEAN | 1.3 | 2 |

Table 1: Words with high and low familiarity ratings 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 second type of human intuitions to be considered 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) 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 8400 stimulus words and is available online.[4]

To collect the data, the subjects were given questionnaires with lists of stimulus words, and were asked to write down for each stimulus word the spontaneous association which first came 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) |

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

### 2.3 Word relatedness

The third type of human intuitions which we consider concerns word relatedness. Landauer & Dumais (1997) introduced a dataset for testing semantic relatedness, namely the synonym portion of the *Test of English as a Foreign Language* (TOEFL). The TOEFL is an often obligatory test for non-native speakers of English who intend to study at a university with English as the teaching language. The data used by Landauer & Dumais had been acquired from the *Educational Testing Service* and comprises 80 test items. As summarized in Rapp (2009), each item consists of a problem word embedded

[3] In the so-called word recognition task test persons are presented strings of characters and their task is to decide whether or not a string matches an English word. It turns out that the average reaction time is inversely related to the familiarity of a word (i.e. the less familiar a word, the longer the reaction time).

[4] http://www.eat.rl.ac.uk/

in a sentence and four alternative words, from which the test taker is asked to choose the one with the most similar meaning to the problem word. For example, given the test sentence *"Both boats and trains are used for transporting the materials*" and the four alternative words *planes*, *ships*, *canoes*, and *railroads*, the subject would be expected to choose the word *ships*, which is supposed to be the one most similar to *boats*.

However, Landauer & Dumais (1997) did not use the test sentences. Instead, only the lists of problem words together with their alternatives were used. A system capable of computing word relatedness should be able to determine for each problem word the alternative word which comes closest in meaning.

Although the TOEFL dataset has been widely used (see e.g. the overview on related work on the ACL Wiki[5]), there are two disadvantages with it: A minor one is that it is not freely available on the web. A more severe one is that it is rather small: This means that statistical variation is strong (which will be illustrated in section 5.3), and that overfitting can easily happen. That is, a system trained on this data may not well perform on other data.

For this reason we decided to come up with a new dataset which avoids these problems. It is based on the index of Fernald's (1896) synonym and antonym dictionary as provided in the Project Gutenberg version.[6] This index lists in alphabetical order English words together with their synonyms. As in the dictionary there is no indication as to the quality of a synonym, in order to avoid arbitrary selections, from this list we removed all words for which several synonyms were listed in the index. In a semi-automatic way, we also removed a number of other items, e.g. those containing multiword units or numbers. As a result, we obtained a list of 4050 words together with their synonyms.

To obtain a dataset analogous to the TOEFL synonym set, we required three alternative words for each item. We could have used random words e.g. taken from the vocabulary of the British National Corpus (BNC). However, as the BNC is from a much later time period, this might have introduced a systematic bias. So we thought we should better use the words from the synonym dictionary itself. Note that the synonyms corresponding to the 4050 words represent a much smaller vocabulary as many of the synonyms are synonyms for several words. For this reason, we used the headwords themselves and applied the following procedure to generate the alternative words from them:

1) We sorted our list of items according to the synonyms in alphabetical order.
2) As the first column of alternative words, we used the given words but shifted them by 1000 positions, i.e. positions 1 to 3050 were matched with 1001 to 4050, and positions 3051 to 4050 where matched with 1 to 1000.
3) Analogous for the second column of alternative words, but here we shifted by 2000 positions.
4) Same for the third column of alternative words, but here we shifted by 3000 positions.

| Word | Synonym | Alternative Words | | |
|---|---|---|---|---|
| abandoned | addicted | rescind | bliss | receipts |
| abdicate | abandon | conflict | indubitable | archaic |
| aberration | insanity | rational | meliorate | assured |
| abetter | accessory | carnal | amicable | urbane |
| abettor | accessory | imbruted | brotherly | policy |
| abhorrence | abomination | kindliness | supposition | resignation |
| abiding | permanent | remain | life | stanch |
| ability | power | chimerical | frontier | diet |
| abject | pitiful | despotic | blanch | fray |
| abjure | abandon | contest | overt | disused |

Table 3: Ten entries from the synonym dataset derived from Fernald (1896).

[5] http://aclweb.org/aclwiki/index.php?title=TOEFL_Synonym_Questions_(State_of_the_art)
[6] http://www.gutenberg.org/files/28900/28900-h/28900-h.htm

To give an impression of the dataset, its alphabetically first ten entries are shown in Table 3. Let us now quickly discuss some properties of the new dataset: The pros are that it is about 50 times larger than the TOEFL dataset and that it can be freely distributed. The cons are that it is based on somewhat outdated language (the dictionary was published in 1896) and that the alternative words were not carefully selected but generated in a somewhat arbitrary fashion. Also, it is not known how test persons would perform on this dataset, whereas for the TOEFL dataset human performance is known at least for some test takers, i.e. non-native speakers of English. A commonality between both datasets is that the synonyms were produced by experts, i.e. reflect the experts' language intuitions.

## 3 Corpora

As in previous work (Rapp, 2014a) 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)[7]
4) ukWaC (British English web corpus of 2 billion words)[8]
5) English Gigaword Corpus 4th edition (4 billion words of newswire text)[9]

Both the MRC familiarity norms and the EAT do not distinguish between uppercase and lowercase characters. For this reason, we also did not make such a distinction and, in a pre-processing step, converted all corpora as well as the human data to lowercase only.

For the results presented later we had to measure the size of our corpora and also of partial corpora. We do this by counting the number of running words. Hereby, to avoid language specific sophistications, we count as a word any string which is delimited by either white space (blanks, tabulator, new line) or by transitions between alpha and non-alpha characters.[10]

## 4 Procedure

### 4.1 Corpus statistics concerning word familiarities (statistics of order zero)

In the case of word familiarities the statistics extracted from the corpora are 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.

| 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 |

Table 4: BNC frequencies and MRC familiarities for the (alphabetically) first ten words covered in the familiarity norms of the MRC database.

[7] We use the English part of the Wikipedia XML Corpus (Denoyer & Gallinary, 2006). Although this is considerably smaller than current versions, it has the advantage that it is an offline copy so that our results can be replicated.
[8] http://wacky.sslmit.unibo.it/doku.php?id=corpora
[9] http://catalog.ldc.upenn.edu/LDC2009T13
[10] Alternatively, it would also be possible to simply count the number of characters for measuring corpus size (though this seems less customary). But word segmentation is required later on anyway (for computing the representativenesss scores).

The two types of data, namely the word familiarities from the MRC database and the word frequencies as extracted from one of the corpora, were merged as exemplified in Table 4 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. Computing corpus representativeness now simply involves taking the logarithm 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 frequency count before taking the logarithm.

### 4.2 Corpus statistics concerning word associations (1st order statistics)

As described in Rapp (2014c), 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 akin to the ones collected from humans is likely to be a good surrogate for everyday language, although – similarly to what we said about word frequencies – word co-occurrence counts constitute only one of many properties of a corpus.

For extracting word associations from corpora, in the literature many algorithms were described (e.g. Wettler & Rapp, 1989; Church & Hanks, 1990; Wettler et al., 2005). In analogy, we used the following procedure: For all words with a BNC corpus frequency of 50 or higher 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 ±10 words around the stimulus word is considered. Hereby the exact distance within the window is not taken into account.

In a further step an association measure was applied to the co-occurrence vectors, namely Ted Dunning's (1993) log-likelihood ratio. 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 5 shows some sample associations as computed using the British National Corpus.

| 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 5: 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 respective primary associative response 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.

### 4.3 Corpus statistics concerning word relatedness (2nd order statistics)

Our algorithm for computing word relatedness consists of the following three steps:

1) Counting word co-occurrences.
2) Applying an association measure to the raw co-occurrence counts.
3) Computing vector similarities.

Steps 1 and 2 are in principle analogous to the previous subsection. Only, as mentioned in Rapp (2009), for computing vector similarities it turns out that it is better to consider a smaller window size (such as ±1 or ±2 around the given word). Also, we used a simpler association measure, namely $log(n_{ij}+1)$, whereby $n_{ij}$ is the number of co-occurrences between words $i$ and $j$, as it slightly outperformed the log-likelihood ratio in this particular setting.

For step 3 (computing vector similarities) we use the standard cosine measure. Table 6 shows some results as obtained using the British National Corpus. For a quantitative evaluation we utilized the TOEFL synonym data as follows: We compared our system's results to the answers as provided in the TOEFL dataset. Remember that in the TOEFL synonym test the subjects had to choose the word most similar to a given stimulus word from a list of four alternatives. Accordingly, in the simulation, we assumed that the system made the right decision if the correct answer was ranked best among the four alternatives. In a further run, we applied exactly the same procedure to the test set derived from Fernald's synonym dictionary.

| | |
|---|---|
| burden | responsibility (0.62), expense (0.61), expenditure (0.59), problem (0.59), cost (0.59) |
| arrogant | rude (0.62), naive (0.61), stupid (0.61), impatient (0.61), haughty (0.61) |
| desperation | panic (0.60), despair (0.60), exasperation (0.59), stillness (0.58), impatience (0.58) |
| memorandum | appendix (0.59), document (0.59), submission (0.57), constitution (0.57), disclosure (0.57) |
| trivial | unimportant (0.63), ridiculous (0.60), trifling (0.60), straightforward (0.60), bizarre (0.60) |

Table 6: Semantic similarities extracted from the BNC for five English words using only vocabulary from the synonym test set based on Fernald (1896).

## 5 Results

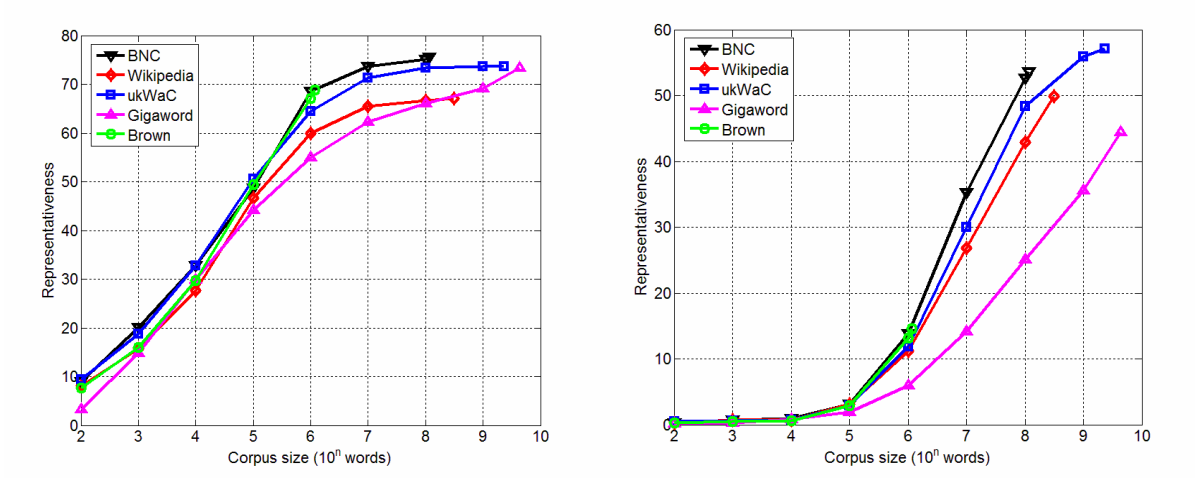
### 5.1 Results based on word familiarities

These results are given in Figure 1a. There we find in graphical form for each of the five corpora the computed Pearson's correlation coefficients between the words' familiarities and their log corpus frequencies. For easier comparison with the other results (which are percentages) we multiply these correlations by 100 and take the product as the familiarity-based *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.

We can see in Fig. 1a 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. The horizontal axis has a logarithmic scale, but still the curves flatten with increasing corpus size, especially above 1 million words.

### 5.2 Results based on word associations

These results are given in Fig. 1b. For each of the five corpora (and their parts) the percentages of primary associative responses are given 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 is between 0 and 100.




(a) Familiarity

(b) Association

(c) Relatedness

(d) Average

Fig. 1: Results for the three approaches and their average.

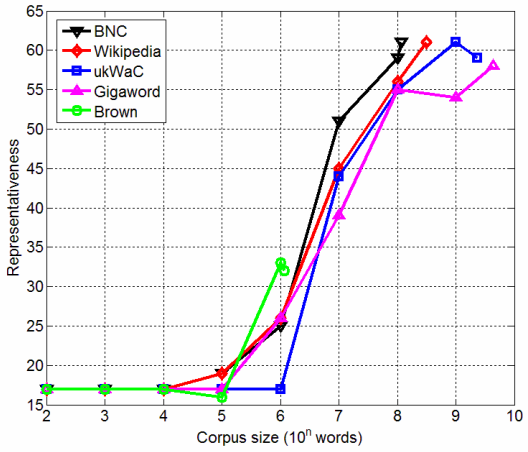



Fig. 2: Results for the TOEFL synonym data.

### 5.3 Results based on word relatedness

The respective results for the TOEFL synonym data (based on a window size of ±1 words) are given in Fig. 2. There we find for each of the five corpora the percentage of TOEFL questions which were answered correctly. These percentages we take as the relatedness-based representativeness of the respective corpus. Note that the level for very small corpora is higher here as in the TOEFL data with a limited number of candidate words there is a better chance to randomly hit the correct word. As can be seen, the curves are somewhat erratic which is an indication that the test set of 80 items is too small.

For this reason, we did not further use these results but replaced them with those from the synonym test set derived from Fernald (1896). The respective results are shown in Figure 1c.[11] As can be seen, the much higher number of test items leads to smoother curves, but nevertheless the tendencies from the TOEFL data are roughly confirmed.

### 5.4 Results based on the overall average

The average of the curves in Fig. 1a to 1c is shown in Fig. 1d. The motivation is that this way all three types of statistics are taken into account in a straightforward way. The underlying reasoning is analogous to the BLEU score (Papineni et al., 2002) used in machine translation evaluation: There n-gram matches between a machine translation and a reference translation are counted separately for n-grams of various lengths, and then the individual scores are combined.

## 6 Discussion

If we compare the curves in Figures 1a, 1b, and 1c it is apparent that the shapes are rather different. This can be explained by the order of the respective statistics: The familiarity-based approach uses statistics of order zero (word frequencies), the association-based approach first order statistics (word co-occurrences), and the relatedness-based approach second order statistics (common context). Although for all three 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 and second order statistics larger corpora would be required to make this happen.

Concerning very small partial corpora, for the familiarity based approach the curves quickly rise, whereas for the association-based and the relatedness-based approaches 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 there is still a chance to find a particular word, but there is almost no chance to find a particular co-occurrence or a common context.

So these discrepancies between the approaches are not a major surprise. Of more interest is a comparison of the results between the different corpora, i.e. their relative performance for each of the methods.

Following Rapp (2014a), concerning the representativeness of our five corpora and their parts, we had 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.

[11] As the Synonym-Dataset involves many very rare words, to reduce data sparseness we used a larger window size of ±2 words to compute these results.

If we compare these hypotheses to the actual results shown in Figures 1a, 1b, and 1c, the findings are as follows:

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

Hypothesis 3 (Brown better than BNC for 1 million words) could sometimes be confirmed but not consistently. Instead, for all approaches the results of these two corpora are fairly close. This indicates that the BNC also seems to have a fairly good balance over the first million words. Concerning the association-based approach, the BNC also has the advantage that its British English should reflect the EAT associations (collected in Edinburgh) better than the American English of the Brown corpus.

Hypothesis 4, namely that ukWaC is better than Wikipedia, is confirmed for the familiarity- and the association-based approach, but not for the relatedness-based approach. Our explanation for the discrepancy is that the relatedness-data contains a larger proportion of outdated and rare words, and that for rare words the coverage of a corpus becomes more important. In this respect, Wikipedia with its wide coverage of topics is likely to have an advantage over the ukWaC corpus.

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

Overall, several of our hypotheses were consistently confirmed by all approaches. This finding provides some evidence that the computed scores are actually related to what might sensibly be considered as the representativeness of a corpus.

Concerning the average representativeness score (Fig. 1d), we can conclude that overall it seems to make sense to balance a corpus, and that corpus heterogeneity is a plus.

## 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, on word associations, and on word relatedness.

Previous work has provided evidence that human word familiarities are based on word frequencies in perceived language (Rapp, 2005), that human word associations are based on the co-occurrences of words (Wettler et al., 2005), and that human relatedness judgments are based on common context (cf. Harris' (1954) distributional hypothesis). Although all of 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, the word associations, and the predictions of word relatedness derived from it resemble these persons' intuitions.

For full and partial versions of five well known English corpora we computed the word familiarities, word associations, and word relatedness scores for test sets of several thousand words. We then, for each corpus, compared the extracted information to the human data, and computed similarity scores which we took as measures of corpus representativeness. We also computed a combined score by averaging the results from all three measures.

A shortcoming of our approach is the following: Our measures are limited in so far as they only consider three particular aspects of corpus representativeness, namely word familiarity word association, and word relatedness. They do not explicitly consider higher level features e.g. concerning syntax, semantics, pragmatics, or style.[12] We nevertheless hope that what we described can serve as a starting point for further discussion.

Concerning future work, a possible strait of research would be to modify the relatedness-based approach in a way that the WordSimilarity-353 Test Collection[13] could be used. This test set provides

[12] In section 5.4, when combining our three approaches, we already mentioned an analogy to the BLEU score. A related commonality is that the BLEU score also has these shortcomings.

[13] http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/

direct similarity estimates between words, so a correlation to corpus-derived estimates could be computed in analogy to what we did for the familiarity-based approach.

We would also like to extend the approach to other corpus statistics which seem relevant for human language processing. For example, we might look at associations when given several stimulus words (see Rapp, 2014b), or we could try to predict a word from its WordNet synset. The latter would have the advantage that WordNets are available for many languages, so the corpus representativeness scores could be measured for a number of languages where other human data is scarce.

Related to this would be the use of the Princeton evocation data[14] which provides human similarity estimates between WordNet synsets. The aim would be to replicate these similarities using multiword associations.

## Acknowledgement

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.

Fernald, J.C. (1896). *English Synonyms and Antonyms*, 19th edition. New York and London: Funk & Wagnalls Company.

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.

Francom, J.; Ussishkin, A. (2011). Converging methodologies: assessing corpus representativeness through psycholinguistic measures. *American Association for Corpus Linguistics.* Georgia State University, Atlanta, GA. http://francojc.files.wordpress.com/2010/01/aacl-2011-converging-methodologies.pdf.

Gries, S.T. (2010). Dispersions and adjusted frequencies in corpora: further explorations. In S.T. Gries, S. Wulff, & M. Davies (eds.): *Corpus Linguistic Applications: Current Studies, New Directions*. Amsterdam: Rodopi, 197–212.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(23), 146–162.

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.

[14] http://wordnet.cs.princeton.edu/downloads.html

Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. (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. (2009**).** The automatic generation of thesauri of related words for English, French, German, and Russian**.** *International Journal of Speech Technology* 11 (3 ), 147–156**.**

Rapp, R. (2011). Language acquisition as the detection, memorization, and reproduction of statistical regularities in perceived language. *Journal of Cognitive Science*, Vol. 12, No. 3, 297–322.

Rapp, R. (2014a). Using word familiarities and word associations to measure corpus representativeness. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland.

Rapp, R. (2014b). Corpus-based computation of reverse associations. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland.

Rapp, R. (2014c). Using word association norms to measure corpus representativeness. In: A. Gelbukh: *Computational Linguistics and Intelligent Text Processing*. 15th International Conference, CICLING 2014, Kathmandu, Nepal. Berlin: Springer. 1–13.

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

Temnikova, I.; Baumgartner Jr., W.A.; Hailu, N.D.; Nikolova, I.; McEnery, T.; Kilgarriff, A.; Angelova, G.; Cohen, K.B. (2014). Sublanguage Corpus Analysis Toolkit: a tool for assessing the representativeness and sublanguage characteristics of corpora. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland.

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.