How specialized are specialized corpora? Behavioral evaluation of corpus representativeness for Maltese

Jerid Francom, Amy LaCross, Adam Ussishkin

Wake Forest University, University of Arizona, University of Arizona
Winston-Salem, NC, USA, Tucson, AZ, USA, Tucson, AZ, USA
francojc@wfu.edu, lacross@email.arizona.edu, ussishki@email.arizona.edu

Abstract

In this paper we bring to light a novel intersection between corpus linguistics and behavioral data that can be employed as an evaluation metric for resources for low-density languages, drawing on well-established psycholinguistic factors. Using the low-density language Maltese as a test case, we highlight the challenges that face researchers developing resources for languages with sparsely available data and identify a key empirical link between corpus and psycholinguistic research as a tool to evaluate corpus resources. Specifically, we compare two robust variables identified in the psycholinguistic literature: word frequency (as measured in a corpus) and word familiarity (as measured in a rating task). We then use three statistical methods to evaluate these comparisons. This research provides a multidisciplinary approach to corpus development and evaluation, in particular for less-resourced languages that lack a wide access to diverse language data.

1. Developing corpus resources for low-density languages

The benefits of developing language corpora that provide readily access to quantifiable language use in naturalistic settings have been widely embraced by many scholars in a diverse set of language disciplines. In addition to the strong tradition of corpus analysis and development in applied research (dictionary (Sinclair, 1987), language teaching (Biber and Conrad, 2001), translation (McEnery and Xiao, 2007) and machine-learning applications), corpus data are increasingly employed in typical theoretical investigation, in particular psycholinguistic studies on language processing (Gilquin and Gries, 2009, for a survey).

However, the great majority of the estimated 5-7,000 languages of the world are 'low-density', i.e. for which robust language resources are limited, or non-existent (Borin, 2009). This fact highlights an obvious lack of empirical coverage of range of possible linguistic diversity – an obstacle for theoretical and applied applications for particular languages and theoretical investigations more generally. To address this gap, many researchers have focused their efforts on developing resources for low-density languages (LDL) (McEnery et al., 2006; Scannell, 2007, inter alia).

Despite best efforts on the part of language researchers, there are unique challenges related to the quality and quantity of available data that researchers must face when developing corpora for LDLs which ultimately may call into question the general applicability of the final product. Whereas access to primary data may be limited both in print and electronic form, creating sometimes insurmountable problems[1], language data that *is* available is often restricted also in terms of its overall representativeness of the target language (i.e. genres/registers, modalities, etc.) (Biber, 1993). Compared to languages such as English language where resources are literally samples of the language, techniques for attaining representativeness for other language corpora are as straight forward given the resources that are available represent almost complete coverage of all language data in existence (Scannell, 2007) – in effect, a representativeness bottleneck.

Under standard evaluation practices many existing projects are considered 'specialized', that is less-than-representative language samples. Accordingly, without some assurance of corpus validity, credible results from low-density language research is limited. However, these smaller and less-diverse language samples do not necessarily misrepresent distributional properties for those linguistic units that have been collected. It is logically possible that some, or all, of the linguistic units contained in the corpus are indeed representative of the larger language body from which it was sampled. Yet the question is, how you know (i.e. determine representativeness)?

In what follows we describe a novel approach to evaluate corpus representativeness that exploits the relationship between corpus linguistics and psycholinguistics.

2. Behavioral data as external validation for corpus resources

Corpus-based evidence is inherently limited in gauging the relative representativeness of a corpus in an absolute sense. A general characteristic of corpus design and evaluation, this limitation is typically addressed by collecting large amounts of data rigorously sampled from a wide variety of sources. For LDL resources, which often lack accessible resources, this is a pressing issue. In this case an external, non-corpus based metric is needed. We propose that evidence from psycholinguistic investigation based on data

We gratefully recognize our colleagues Dr. Albert Gatt (U of Malta) and Jeff Berry (U of Arizona) for their invaluable assistance and acknowledge funding from the United States National Science Foundation (BCS-0715500) to Adam Ussishkin.

[1] Difficulties in attaining data do not always stem from the number of speakers of language, but may in fact reflect the interaction of various extra-linguistic factors (cultural, economic, political, etc.).

garnered from representative linguistic corpora provides a potential external source for such a metric.

2.1. Frequency predicts behavior

Relative frequency differences between linguistic items extracted from language corpora has been demonstrated as the one of the most reliable variables in psycholinguistics. 'Frequency effects' have been shown to robustly affect language processing at various levels (lexical, syntactic, etc.) and are detectable in multiple psychological tasks and measures; including language comprehension accuracy, sentence reading times, word naming times, lexical decision and word familiarity latencies, etc., where over and over, (log) frequency predicts behavior. In addition, this variable has been studied in a number of well-documented languages for which there exist representative corpora, including English (Grainger, 1990), Dutch (Grainger, 1990), German (Penke and Krause, 2002), and Spanish (Alvarez et al., 2001), among others. These studies exploit existing rich corpora drawn from accepted representative samples, such as Francis and Kucera ((Francis et al., 1982); based on the 1.1 million word Brown Corpus), CELEX ((Baayen et al., 1993); based on the COBUILD project, with 17.9 million words, 42.4 million Dutch words, and 5.4 million German words), and the Spanish Word Pool (Alameda and Cuetos, 1995). So robust is the effect that controlling for frequency is standard practice in psycholinguistic research –even in cases in which frequency itself is not a treatment variable.[2]
A related measure key to the current investigation is word familiarity, which indexes how familiar a word is among speakers of a language community. Empirically, word familiarity is closely tied to word frequency, albeit indirectly given that word familiarity is a subjective measure. However, an important argument in favor of word familiarity as an index of frequency over other tasks (e.g. lexical decision) is that it reflects both visual and auditory frequency (Gernsbacher, 1984; Connine et al., 1990).[3]
In short, the strong relationship between these two variables, word frequency on the one hand, and word familiarity on the other, sets the stage for an investigation comparing the two. In particular, we can study how close this relationship is for a particular language for which there exists a corpus (which can provide word frequency measurements) and for which there exists a population of native speakers (who can provide word familiarity measurements).

2.2. Behavior can predict expected frequency

We propose here that the robustness of frequency effects highlighted in behavioral evidence mounting since the 60s can be harnessed to provide external evaluation for corpus resources. That is, given the strong evidence that relative frequency of linguistic units from representative language corpora predicts language behavior in a number of tasks, and for a wide variety of languages we suggest that, inversely, language behavior can be harnessed as indirect evidence of the goodness of fit of the distribution of linguistic units in a corpus.
Behavioral evaluation of language collections provides a needed non-corpus based metric for gauging the external validity of language samples, generally, and for LDLs, this validation presents an opportunity to address pressing issues concerning the nature of 'specialized' language resources. Short-term assessment of language resources is essential in order for researchers to gauge the viability of their resources.
Exploiting the correlation between frequency counts and behavioral data is not completely novel, however. In a recent set of studies dealing with the dispersion of linguistic units in a corpus[4] (Gries, 2008; Gries, 2009) employes evidence from language processing to provide an external criterion to guide selection between competing frequency calculation metrics. Gries argues that while information about the observed frequency of a linguistic item is descriptively important, it is also psychologically relevant as demonstrated in a number of language disciplines. This connection between psychological data and corpus data provides an important external reference point when theory-internal metrics cannot serve as the absolute criterion for validating frequency measures.
The current paper aims to develop complementary line of inquiry to address corpus representativeness. In the following section, we turn to the results of a series of experiments aimed at assessing a recently developed LDL resource for Maltese.

3. A case study from Maltese

The corpus resource to be evaluated is a lexical corpus created for Maltese (Francom et al., 2009). The PsyCoL Maltese Lexical Corpus (PMLC) was developed by the PsyCoL lab at the University of Arizona and can be accessed online at http://psycol.sbs.arizona.edu/resources/. It contains 3,323,325 total tokens (53,000 which are unique, for a Token/type ratio of 1.6%) which represent two distinct web crawling efforts; one by the PsyCoL lab (59.8%) and the other by Dr. Albert Gatt (40.2%) at U of Aberdeen/ U of Malta.
This data represents the largest collection of accessible Maltese language data but is potentially limited in terms of representativeness as the seed sources for all web crawls were online newspapers.[5] Given the low-density nature of Maltese, obtaining a wider, more widely varied sample sample was not feasible. In this way, the PMLC resource is endemic of many resources for low-density languages and as such a prime candidate to explore the intersection between corpus representativeness and psycholinguistic research through frequency effects.

[2]For languages without such resources, it remains an open question whether or not frequency behaves in the same way. It seems reasonable to assume that it does, but this assumption should not preclude resource creation and psycholinguistic investigation for LDLs.

[3]Other supporting evidence for the robust nature of subjective word familiarity can be found in (Balota et al., 2001) and (Nusbaum et al., 1984).

[4]The number of occurrences of any given item across documents, within documents, etc; not only the raw number of linguistic items contained in the corpus

[5]Those sources include Illum, L-orizzont, Kullħadd, In-Nazzjon and Leħen is-Sewwa

3.1. Hypothesis

In this section we explore the hypothesis that robust predictors of linguistic behavior can serve as an external metric of corpus representativeness. As a first step towards this goal we compared results from an independently conducted behavioral experiment on the word familiarity ratings for Maltese to frequency data for those words that appear in the PMLC.

The empirical focus of our comparison between corpus data and behavioral data for Maltese is Semitic-origin verbs. As a hybrid language, Maltese contains two types of verbs (Borg and Azzopardi-Alexander, 1997; Mifsud, 1995): Semitic-origin or Arabic-style verbs, and loan verbs. Here, we limit our inquiry to the Semitic-origin verbs for the simple reason that they follow a very clear set of restrictions.

In short, on the assumption that word frequency is correlated with word familiarity ratings, we predict that subjective word familiarity ratings provided by native speakers should approximate or match verb frequencies in the PMLC if the corpus is representative (for this linguistic dimension). In the following section, we lay out the methodological details on how the evaluation was conducted.

3.2. Methods

The purpose of the word familiarity experiment was to determine subjective word familiarity ratings for a subset of the Maltese vocabulary. In particular, the items in the experiment were limited to all 1536 Semitic-origin verbs of Maltese, as contained in the Aquilina (2000) Maltese-English dictionary; considered to be the authoritative word list of Maltese. The experiment involved 107 participants, all of whom were native speakers of Maltese, and who logged in to a secure website to participate in the experiment. In each trial, each participant responded to one of the 1536 items by rating how familiar they were with the item. This was accomplished using a slider bar, whose left edge corresponded to "this item is not familiar to me" and whose right edge corresponded to "this item is very familiar to me." Because of the large number of total experiment items, each participant was given the option of exiting the experiment at any point, and all unrated items for a participant were batched at the top of the randomized list of items given to the following participant. On average, each verb was rated 6.01 times, and each subject on average provided a word familiarity rating for 55.15 verbs. To analyze the data, each rating was converted to a percentage scale with a range from 0 to 100.

In addition to the behavioral experiment, we also used the PMLC to measure word frequency. Using this corpus, we calculated word frequency for the subset of Semitic verbs employed in the word familiarity task that also appear in the corpus. Using regular expressions to extract the verb patterns, we searched the corpus for all inflected forms of each verb, which was then coded for frequency expressed as a logarithmic dependent measure; returning a total of 447 verbs.

3.3. Analysis

In order to provide a broad assessment of the correlation between the two measurements we obtained (word familiarity ratings and word frequency in the PMLC) we analyzed the data using three methods. First, we ran a statistical regression to determine whether at the level of individual tokens there exists any correlation between a given words familiarity rating and its frequency. Second, we binned our dependent measures into different groups to see if any correlation could be found between familiarity and frequency based on these groups. Finally, taking advantage of the classical Semitic binyan patterns inherent in Maltese verbs, we compared these binyan patterns both in terms of familiarity and frequency to test whether any of the same relationships obtained between binyanim for these two dependent measures.

3.3.1. Verb frequency

The most direct comparison between familiarity ratings and frequency counts is an assessment of token correlation. The graphic in Figure 1 shows a trend towards more frequent verbs to also be judged as more familiar. Statistically, word familiarity ratings and verb frequency (log) show a weak correlation ($r = .14$). On the surface, this low correlation suggests that familiarity ratings do not predict verbal frequency in the PMLC – despite the visual pattern to the contrary.

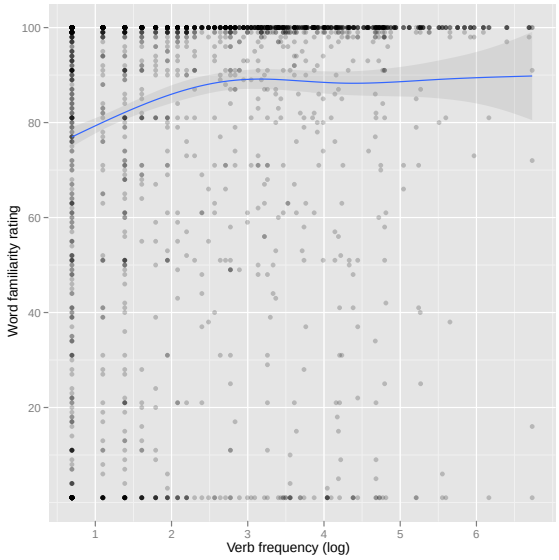

Figure 1: Correlation between word familiarity and word frequency

Looking at the distribution of participants' responses and frequency counts for verbs analyzed in this sample (seen in Figure 2 and 3), there are two issues related to participant's responses that are important to note: 1) responses are skewed towards higher ratings more generally, and b) responses showed a tendency towards a categorical distribution (that is, participants did not fully exploit the wide range of responses provided to them.[6]) which on the one hand compresses the response scale making differences more

[6]Magnitude Estimation was used as the elicitation technique in this rating task. Recent literature suggests that participants may in fact tend towards categorical distributions even in spite of a more diverse set of response options (Sprouse, 2009).

difficult to detect and on the other, increases the potential for more variability associated with each verb.

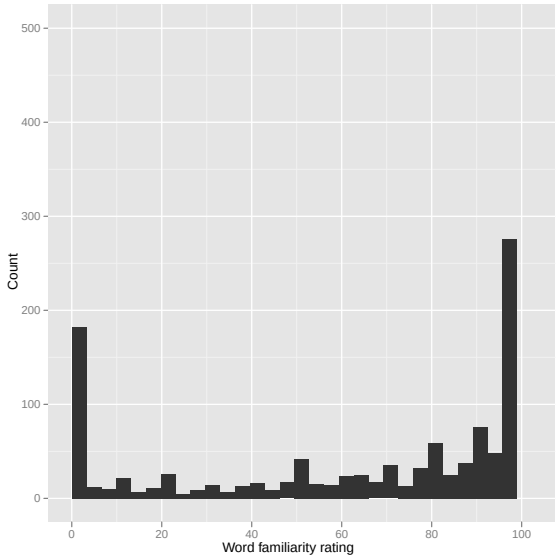

Figure 2: Distribution of word familiarity ratings

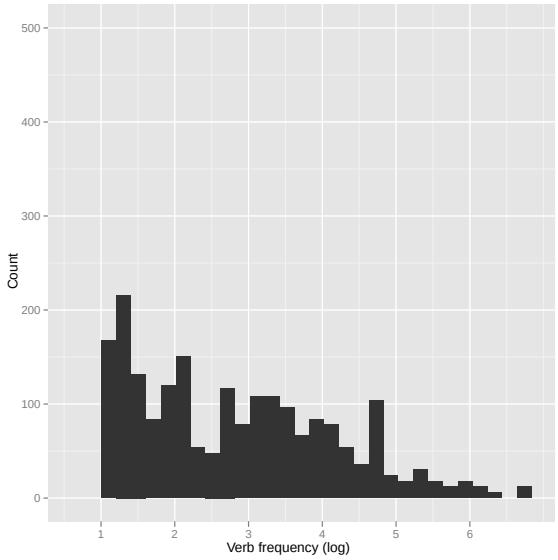

Figure 3: Histogram of verb frequency counts

Considering the difficulties in interpreting these results, we decided to group verbs by frequency intervals to approach the relationship more generally. These interval bins were calculated by cutting the log frequency distribution in two, three, and five equal bands.

Mean word familiarity ratings were calculated for each interval, the results are shown in Tables 1, 2, and 3. Observationally, two-bin, three-bin and five-bin groupings show the expected trend towards higher familiarity ratings for higher frequency verbs overall. Yet, the five-bin grouping diverges from this trend in the Mid and Mid-low intervals (roughly between the log frequency values 3 and 5).

Subsequently, stepwise comparisons between frequency intervals were analyzed statistically with a Linear Mixed-effect model with word familiarity rating as the dependent variable, frequency interval as the fixed-effect and verb and subject as random effect variables. All contrasts for two-bin intervals (High/Low $\beta = 4.2$, $t = 2.0$) and three-bin intervals (High/Mid $\beta = 7.1$, $t = 3.9$; Mid/Low $\beta = 7.0$, $t = 2.2$) were significant as was the Low/Mid-low ($\beta = 10.3$, $t = 4.9$) and the pairwise comparison for the polar intervals (High/Low $\beta = 7.1$, $t = 3.9$) in the five-bin grouping.

In sum, the results suggest that there is a trend along the lines predicted by the current hypothesis.

Word frequency	Mean word familiarity
Low	83.18
High	88.39

Table 1: Word familiarity ratings grouped by two frequency intervals

Word frequency	Mean Word familiarity
Low	81.30
Mid	88.61
High	89.63

Table 2: Word familiarity ratings grouped by three frequency intervals

3.3.2. Binyan

The third approach taken in our investigation took advantage of the structure of the Maltese verbal system, which for the Semitic-origin verbs that we tested in the word familiarity experiment, is organized into a set of categories (known in the Semitic literature as binyanim), each of which is identifiable based on both its morphosyntactic/semantic and its prosodic properties. In Maltese, there exist a total of 9 binyanim, numbered from 1-10 (binyan 4 has been lost through diachronic change, and doesnt exist in Maltese). The numbering system used to refer to the binyan system corresponds to the numbers used in the traditional Arabic grammars for the analogous categories. As in other Semitic languages, Maltese verbs can be either strong (with 3 or more consonants) or weak (with fewer than three consonants).

In our analyses below, since our focus here is the binyan system, we conflate strong and weak verbs, though we recognize that future research may need to take this distinction into account. Because each binyan is easily identifiable based on its prosody, we were able to use the PMLC to perform a frequency calculation of every Maltese verb binned by binyan.

Our hypothesis that frequency might matter for Maltese is based on earlier evidence from the related Semitic language Hebrew (Moscoso del Prado Martín et al., 2005; Ussishkin et al., in progress), which shows a clear effect of word frequency in both the visual and auditory modalities. Ussishkin et al. (in progress) also identify a binyan size effect for Hebrew, providing further justification for our Maltese

Word frequency	Mean Word familiarity
Low	78.96
Mid-low	89.14
Mid	88.22
Mid-high	88.66
High	89.56

Table 3: Word familiarity ratings grouped by five frequency intervals

study here.

Initial inspection of the distribution of our rating scores by binyanim revealed that some binyanim contain a very small number of verbs. These binyanim (3,6,8,9,10) were dropped from subsequent analyses as each of these binyanim were not sufficiently represented in order to perform reliable a statistical evaluation. Figure 4 illustrates both familiarity ratings by binyan as well as the fact that five binyanim are sparsely populated.

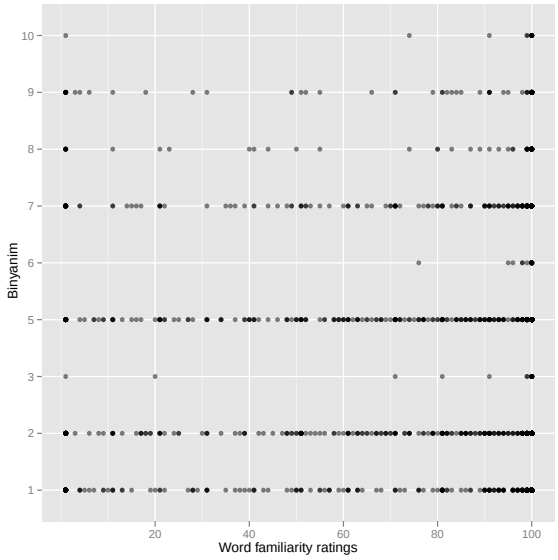

Figure 4: Number of word familiarity ratings by binyanim

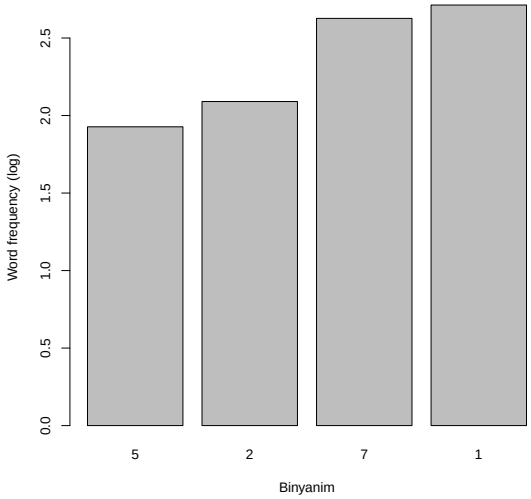

Figure 5: Binyanim by word frequency

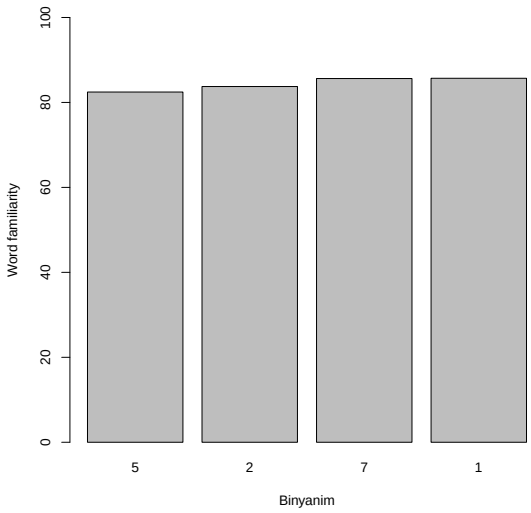

Figure 6: Binyanim by word familiarity

Our goal was to then calculate mean frequency for binyan and test for significant differences in the frequency of each binyan as a category. Frequency was calculated using natural logarithm, as in previous analyses. The results of our frequency calculations are given visually in Figures 5 and 6. With respect to word frequency differences, significant contrasts were found for Binyan 7 and Binyan 2 ($\beta = .54$, $t = 6.0$) on the one hand, and Binyan 7 and Binyan 5 (2-5 $\beta = 1.15$, $t = -2.2$) on the other hand.

For word familiarity, no statistical difference was found for familiarity ratings in any of the pairwise comparisons between the different Binyanim – despite the fact that both frequency rankings and rating rankings match in order. This result may reflect methodological problems related to ceiling effects mentioned in the verb frequency grouping analysis. An unavoidable feature of the current investigation, it may be fruitful in future investigations to run the familiarity rating experiment using a 7-point scale, rather than asking participants to rate familiarity using a slider in order to gauge the effect of response scale on rating distributions. In any case, the fact that the word familiarity results show no categorical effects of the binyan system is potentially of interest in and of itself, or may obscure a correlation between binyan frequency and language processing. We are encouraged, though, that the corpus does appear to reflect categorical differences between several of the verbal binyanim of Maltese, which is in line with other evidence reported in the literature, and that relative rankings suggest a link between both measures.

3.4. General assessment

We have reviewed the evidence for the connection between frequency counts extracted from representative samples of language and behavior in psychological experimentation

and proposed that this very link can be harnessed to predict expected frequency distributions in language corpora. A first pass in the exploration of this hypothesis was provided by evaluating the correlation between data garnered from a recent word familiarity rating experiment, in which native Maltese speakers rated the subjective familiarity of verbs, and corresponding frequency counts from a recently created Maltese lexical corpus. Findings here suggest that the PMLC shows encouraging distributional patterns in broad terms. Data supporting this position comes from significant rating contrasts between verb frequency intervals in two-bin and three-bin groupings. In addition, a pattern towards higher frequency verbs to be rated as more familiar appears across token, bin and binyan approaches pointing to a general correlation.

However, drawing strong conclusions from these results would be misguided. Although on the right track, a number of key contrasts are not found that would provide more definitive connections between frequency and rating scores. There are a number of probable reasons why significant contrasts are not found, despite the global trend found in these analyses. The most obvious is that the corpus is indeed 'specialized' and not representative, at least not along the dimensions we have explored; a plausible conclusion. A second is that the general hypothesis that the finding that 'behavior is predicted by frequency' cannot be applied by analogy to 'frequency is predicted by behavior'. However, evidence for frequency effects is compelling and robust across languages, tasks and linguistic variables, downplaying chances that the hypothesis is entirely false. Instead, we believe that there is reason to believe that the major shortcoming in these analyses lies not with the corpus, *per se*, but rather with the distribution of rating scores in the behavioral experiment. As mentioned, the categorical distribution and skew towards higher ratings undermine the power of the statistical analyses to provide reliable evaluations. Furthermore, a bid of confidence for the PMLC comes from a corpus-internal correlation of word length and word frequency, as seen in Figure 7. The data roughly conforms to distributional patterns of word length and word frequency related to Zipf's Law (Zipf, 1949; Li, 1992); specifically, word length is inversely correlated with word frequency. In this way, the PMLC sample demonstrates an expected distribution of a strong cross-linguistic pattern (Bates et al., 2003).

4. Conclusion

In this paper we have proposed a novel methodology for evaluating corpus resources. This approach exploits a long-standing, but typically one-way, connection between corpus linguistics and psycholinguistics. We have highlighted one of the most robust findings psycholinguistics: the frequency of linguistic units predicts language behavior. Our attempt here has been to reverse the logic and evaluate to what extent language behavior can predict the frequency of linguistic units in a corpus. Such a prediction provides a possible angle to provide external validation for resources, such as those for low-density languages, that cannot feasibly acquire more data (in the short-term) as a solution to representativeness.

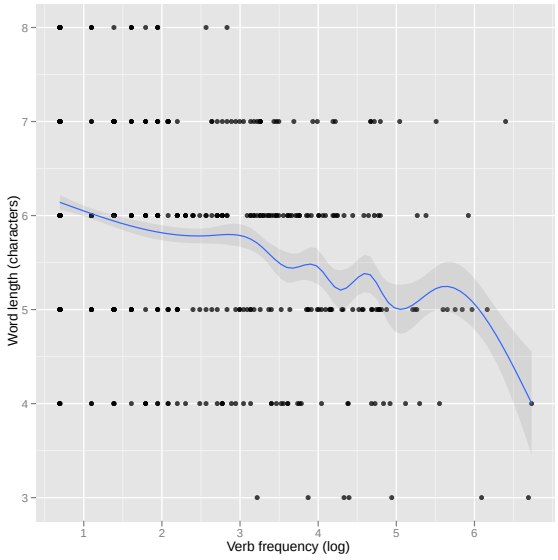

Figure 7: Correlation between word length and word frequency

Results provided here from Maltese suggest that there is a general correlation between word familiarity ratings from native speakers and observed frequency counts in the Maltese corpus – despite distribution irregularities in the rating scores collected from in the word familiarity experiment. More importantly, this work sheds light on ways to create and test corpora for low-density languages in particular, and other larger, more representative samples generally, using a combination of existing methodologies and paradigms. In this way, we believe this approach encourages cross-discipline approaches to resource development and theoretical investigation.

5. References

J.R. Alameda and F. Cuetos. 1995. Diccionario de frecuencias de las unidades lingüísticas del castellano.

C.J. Alvarez, M. Carreiras, and M. Taft. 2001. Syllables and morphemes: Contrasting frequency effects in Spanish. *Journal of Experimental Psychology Learning Memory and Cognition*, 27(2):545–555.

J. Aquilina. 2000. *Maltese-English Dictionary*. Midsea Books, Santa Venera, Malta.

R.H. Baayen, R. Piepenbrock, and H. van Rijn. 1993. The CELEX lexical database (CD-ROM). *Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA*.

D.A. Balota, M. Pilotti, and M.J. Cortese. 2001. Subjective frequency estimates for 2,938 monosyllabic words. *Memory and Cognition*, 29(4):639–647.

E. Bates, S. D Amico, T. Jacobsen, A. Szekely, E. Andonova, A. Devescovi, D. Herron, C.C. Lu, T. Pechmann, C. Pléh, et al. 2003. Timed picture naming in seven languages. *Psychonomic Bulletin and Review*, 10(2):344–380.

D. Biber and S. Conrad. 2001. Quantitative corpus-based research: Much more than bean counting. *TESOL Quarterly*, pages 331–336.

D. Biber. 1993. Representativeness in corpus design. *Literary and linguistic computing*, 8(4):243–257.

A.J. Borg and M. Azzopardi-Alexander. 1997. *Maltese*. Routledge.

L. Borin. 2009. Linguistic diversity in the information society. *Proceedings of the SALTMIL 2009 workshop on Information Retrieval and Information Extraction for Less Resourced Languages.*, pages 1–7.

C.M. Connine, J. Mullennix, E. Shernoff, and J. Yelen. 1990. Word familiarity and frequency in visual and auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(6):1084–1096.

W.N. Francis, H. Kučera, and A.W. Mackie. 1982. *Frequency analysis of English usage: Lexicon and grammar*. Houghton Mifflin Harcourt (HMH).

J. Francom, A. Ussishkin, and D. Woudstra. 2009. Creating a Web-based Lexical Corpus and Information-extraction Tools for the Semitic Language Maltese. *Proceedings of the SALTMIL 2009 workshop on Information Retrieval and Information Extraction for Less Resourced Languages.*, pages 9–16.

M.A. Gernsbacher. 1984. Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, 113(2):256–281.

G. Gilquin and S.T. Gries. 2009. Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory*, 5(1):1–26.

J. Grainger. 1990. Word frequency and neighborhood frequency effects in lexical decision and naming. *Journal of Memory and Language*, 29(2):228–244.

S.T. Gries. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, 13(4):403–437.

S.T. Gries. 2009. Dispersions and adjusted frequencies in corpora: further explorations.

W. Li. 1992. Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, 38(6):1842–1845.

A. McEnery and R. Xiao. 2007. Parallel and comparable corpora: What are they up to?

T. McEnery, R. Xiao, and Y. Tono. 2006. *Corpus-based language studies: an advanced resource book*. Routledge.

M. Mifsud. 1995. *Loan verbs in Maltese: a descriptive and comparative study*. Brill.

F. Moscoso del Prado Martín, A. Deutsch, R. Frost, R. Schreuder, N.H. De Jong, and R.H. Baayen. 2005. Changing places: A cross-language perspective on frequency and family size in Dutch and Hebrew. *Journal of Memory and Language*, 53(4):496–512.

H.C. Nusbaum, D.B. Pisoni, and C.K. Davis. 1984. Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words. *Research on Speech Perception Progress Report*, 10:357–376.

M. Penke and M. Krause. 2002. German noun plurals: A challenge to the dual-mechanism model. *Brain and language*, 81(1-3):303–311.

K. Scannell. 2007. The Crúbadán Project: Corpus building for under-resourced languages. *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop, incorporating Cleaneval*, page 5.

J.M. Sinclair. 1987. *Looking up: an account of the COBUILD project in lexical computing and the development of the Collins COBUILD English language dictionary*. Collins ELT.

J. Sprouse. 2009. Revisiting Satiation: Evidence for an equalization response strategy. *Linguistic Inquiry*, 40(1).

A. Ussishkin, J. Berry, A. LaCross, H. Velan, and A. Twist. in progress. Family size in Hebrew auditory word recognition. Ms., University of Arizona and Hebrew University of Jerusalem.

G.K. Zipf. 1949. Human Behavior and the Principle of Least Effort.