# Tracks in the Mind: Differential Entrenchment of Common and Rare Liturgical and Everyday Multiword Phrases in Religious and Secular Hebrew Speakers

**Jonathan Berant (jonatha6@post.tau.ac.il)**
School of Computer Science, Tel-Aviv University
Tel-Aviv, 69978, Israel

**Catherine Caldwell-Harris (charris@bu.edu)**
Department of Psychology, Boston University
64 Cummington St., Boston, MA 02215, USA

**Shimon Edelman (se37@cornell.edu)**
Department of Psychology, Cornell University
Ithaca, NY 14853, USA

**Abstract**

We tested the hypothesis that more frequent exposure to multi-word phrases results in deeper entrenchment of their representations, by examining the performance of subjects of different religiosity in the recognition of briefly presented liturgical and secular phrases drawn from several frequency classes. Three of the sources were prayer texts that religious Jews are required to recite on a daily, weekly, and annual basis, respectively; two others were common and rare expressions encountered in the general secular Israeli culture. As expected, linear dependence of recognition score on frequency was found for the religious subjects (being most pronounced for men, who are usually more observant than women); both religious and secular subjects performed better on common than on rare general culture items. Our results support the notion of graded entrenchment introduced by Langacker and shared by several cognitive linguistic theories of language comprehension and production.

Keywords: entrenchment; prefabs; frequency effects; sentence processing; multiword expressions; collocations.

## Introduction

In processing an incoming utterance or in generating a new one, a user of language has the choice of two computational strategies. Using the terminology of Sinclair (1991), these strategies involve, respectively, (i) the *open-choice* principle, according to which units are looked up in the lexicon and then composed together according to the rules of grammar, and (ii) the *idiom* principle, according to which a pre-fabricated version of the utterance, if one exists, is drawn from an extended phrasal lexicon (Becker, 1975), rather than being constructed all over again "on the fly." Balancing these two options in language corresponds in more general terms to the familiar computational trade-off between processing time and memory capacity (Yang, 2005).

One of the foundational insights of cognitive linguistics is that these options need not be entirely distinct. As noted by Langacker (1987, p.59), "Linguistic structures are more realistically conceived as falling along a continuous scale of **entrenchment** in cognitive organization. Every use of a structure has a positive impact on its degree of entrenchment, whereas extended periods of disuse have a negative impact. With repeated use, a novel structure becomes progressively entrenched, to the point of becoming a unit; moreover, units are variably entrenched depending on the frequency of their occurrence."

In the two decades since the publication of Langacker's volume, the concept of entrenchment became a central part of the item-based theories of language acquisition (Tomasello, 2003; Goldberg, 2005). Although empirical evidence for entrenchment in L1 and L2 acquisition has been accumulating steadily (e.g., Theakston, 2004; Eskildsen & Cadierno, 2007), many of the originally posited characteristics of this phenomenon are still waiting to be thoroughly explored. In the present paper, we focus on one such characteristic: the graded dependence of entrenchment on exposure.

Harris (1998) demonstrated that the last words of four-word English idioms (such as "great minds think alike") are processed more quickly when they are followed by either the first two words or the middle two words of the idiom, suggesting that entrenched idiom-level representations exist. She was also able to show that common word pairs (such as "tax bill"), but not random ones (such as "tag bill"), activate unitized representations for the pair, as well as schematic representations that are consistent with its syntactic structure.

More recently, Caldwell-Harris and Morris (submitted) reported a correlation between the purported degree of entrenchment of word pairs (as estimated by their Google frequencies) and the probability of the subjects perceiving them in the "canonical" order when presented with the reversed order (e.g., seeing "zip code" when presented with "code zip").

Although it is now possible to use corpus tools and Google searches to estimate the frequency of a given multi-word expression in various contexts, neither expressions specifically described as idioms nor merely common multi-word expressions offer the possibility of *controlling* the subjects' exposure to the stimuli. Without such control, the frequencies observed in a corpus may be only loosely related to each subject's experience, making it difficult to

demonstrate the psychological reality of graded entrenchment.

In the study reported here, we circumvented this difficulty by resorting to stimuli for which the frequency of exposure could be estimated for a particular subject population. Thus, we predicted that observant Israeli Jews would exhibit differential entrenchment for Hebrew multi-word expressions taken from liturgical texts that are recited on different occasions during the daily, weekly, and annual prayer cycles. A control group of subjects comprised secular Israelis. The two groups were tested both with these religious expressions and with rare and common phrases that are part of the everyday Israeli cultural experience.

## Praying Customs in Orthodox Judaism

Observant Orthodox Jews are required by their religion to recite three prayers every day. On Saturdays, which is the Jewish holy day of the week, the prayers are longer and the service is substantially different. The prayers recited over the annual High Holy Days are further prolonged and modified. Prayers can be recited in privacy, without a quorum, or preferably in a synagogue, where the service is longer. The High Holy Days service carries a higher emotional charge than the more frequent mundane prayers.

The majority of the Jewish population in Israel is secular. Generally, secular people do not pray at all, although a substantial proportion does attend some of the services in the High Holy Days. In the religious sector, the variability in praying customs is high: some people attend all services mentioned above in a synagogue, others pray daily in private, and some only pray on a weekly basis or even less often. Furthermore, within the religious sector, there are pronounced cultural differences between the sexes: traditionally, men are expected to be more meticulous about praying than women.

Thus, there are two factors that affect the frequency with which an individual is exposed to phrases from the Jewish liturgical texts. The first is the objective frequency of the phrase in Jewish prayers. Phrases that appear in daily prayers are presumably recited more often than phrases that appear in weekly prayers, which in turn are recited more often than phrases that appear only in annual prayers. The second factor is the specific praying habits of the individual. The more observant a person is, the more we expect him or her to have been exposed to religious phrases. We can use these two factors to predict the level of entrenchment of a phrase from a prayer in a religious individual. This entrenchment is the product of being naturally exposed to liturgical Jewish phrases throughout many years of experience.

Table 1: Examples of phrases.

| Phrase Type | Example |
|---|---|
| Daily | *morid hatal* |
| Weekly | *nafshi yeshovev* |
| Annual | *bnei maron* |
| Mixed | *zore'a ha'amim* |
| Common | *shalom xaver* |
| Rare | *divrey rahav* |

## Methods

### Participants

Fifty-one native Hebrew speakers, ranging in age from 18 to 36, were categorized according to whether they self-reported as being religious or secular; 19 participants attested that they were secular (11 females and 8 males), and the other 32 attested that they were religious (19 females and 13 males). Each participant completed a questionnaire detailing praying habits (frequency of praying and whether in private or at synagogue).

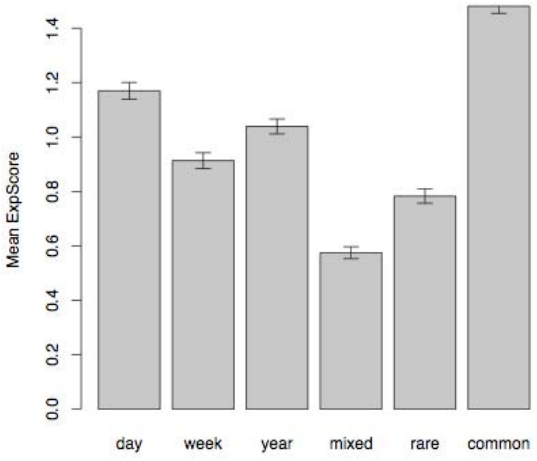


Figure 1: Mean experiment Score (ExpScore) and standard error across levels of frequency for all of the subjects.

### Materials

Six groups of stimuli were constructed (see examples in Table 1). Three of these were religious phrases, categorized according to frequency of recitation (daily, weekly, and annual). Two groups were nonreligious phrases, categorized as common and rare. The sixth group was constructed out of words that appear separately from each other in Jewish prayers and do not form cohesive phrases (mixed). Each phrase group comprised eleven 2-word phrases and four 3-word phrases. The length in characters of phrases was similar across all categories (mean=12.1, std=0.55). Religious phrases were extracted from Jewish prayers that are typically[1] recited daily, weekly, or yearly. Phrases were selected to have comparable semantic and syntactic complexity. The common phrases were drawn from Israeli culture and included political slogans, names of famous TV

[1] A person may also conceivably encounter these phrases outside of the prayer routine, e.g., in the course of studying or of some other activity.

shows, and popular songs. The rare phrases were selected from modern Hebrew literature and poetry. Google counts confirmed that the phrases in the rare group were significantly less common than phrases in the common group (mean log frequency of rare phrases = 1.6, common phrases = 4.6, $p<0.0001$).

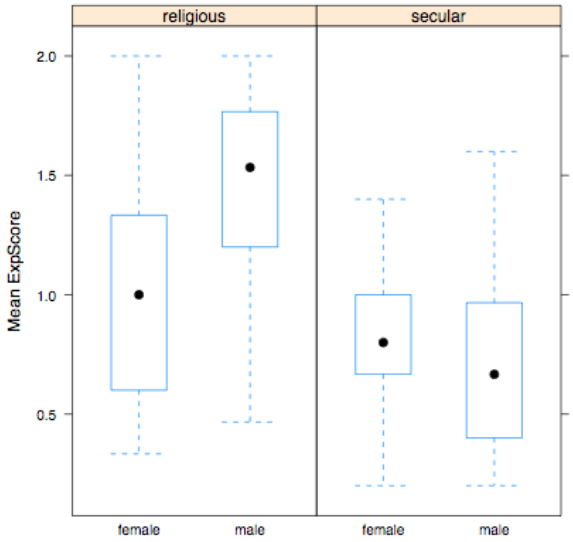


Figure 2: Mean ExpScore per subject per frequency boxplot on religious phrases across the Religion and Sex factors.

## Procedures

In the perceptual identification task, phrases were briefly displayed in the center of the screen followed by the masking stimulus ('&&&&&&'). Participants immediately verbally reported the words to the experimenter who recorded their response, and then participants initiated the subsequent trial with a key press. Exposure durations were set individually for each participant, based on performance in practice trials. For two-word phrases, exposure duration ranged from 50-80 *ms* across the participants (mean exposure 71.71 *ms*); for 3-word phrases, durations were set to be 20 *ms* longer.

Stimuli were displayed using the DirectRT software (www.empirisoft.com), under Windows XP installed on a 2.16 GHz Intel 2 Core Duo MacBook. The perceptual identification task was explained to the participants both orally and in a written form prior to the practice trials. The 90 experimental trials were randomly ordered. Responses were scored for accuracy using a three-point scale: 0 for recognition of less than one word, 1 for correct recognition of at least one word in the phrase, 2 for exact and complete recognition of the phrase.

After completion of the perceptual identification task participants were tested on their knowledge of Jewish prayer texts in a paper-and-pencil test. The subject was asked to complete 17 phrases taken from various Jewish prayers. (E.g., *barux shem kvod* ____). Zero points were given for a phrase left incomplete, one point for partial completion, and two points for perfect completion of the phrase.

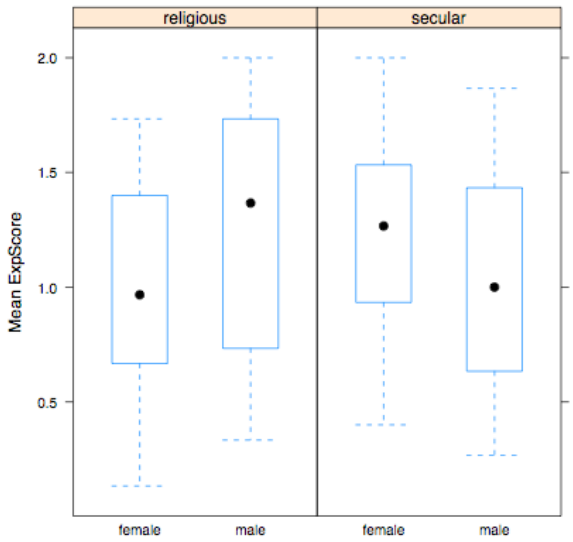


Figure 3: Mean ExpScore per subject per frequency boxplot on secular phrases across the Religion and Sex factors.

## Results

The analysis examined the relationships between the dependent variable, ExpScore (the perceptual identification score, measured on a 0,1,2 scale) and five predictors: Religion (a 2-level factor, secular vs. religious); Sex (a 2-level factor, male vs. female); Frequency (a 6-level factor, daily, weekly, annual, mixed, common and rare); PencilScore (average score the subject got in the paper-and-pencil test, 0-1 scale); and PrayersPerDay (number of prayers recited by the subject every day, with possible values being 0,0.5,1,2,3). In addition to those fixed effects, two random factors were included in the analysis: Subject and Item.

Because items from the 'mixed' condition were not true phrases, they were dropped from the analysis after ascertaining that their score was significantly lower than that in any other condition (see Figure 1). Data were then split into two subsets, each with an ordered Frequency factor: religious phrases (daily, weekly, and annual) and secular phrases (rare vs. common).

Analysis was carried out in two stages, using the R processing environment (www.r-project.org). First, repeated-measures ANOVA tests were performed after averaging the scores separately by subjects and by items (Clark, 1973). Because effects that turn out significant in separate by-subject and by-item analyses may still be unreliable when all the random factors are considered jointly (Raaijmakers et al., 1999), we also fit a series of mixed linear models to the data using the lme4 package (Baayen, in press), which can handle multiple crossed random effects. Because high correlations were found

between the Religion, PencilScore and PrayersPerDay variables (PencilScore-PrayersPerDay: r=0.78, p<0.0001, Religion-PencilScore: r=0.92, p<0.0001, Religion-PrayersPerDay: r=0.68, p<0.0001), separate models were fit for each of these variables as a stand-in for religiosity.

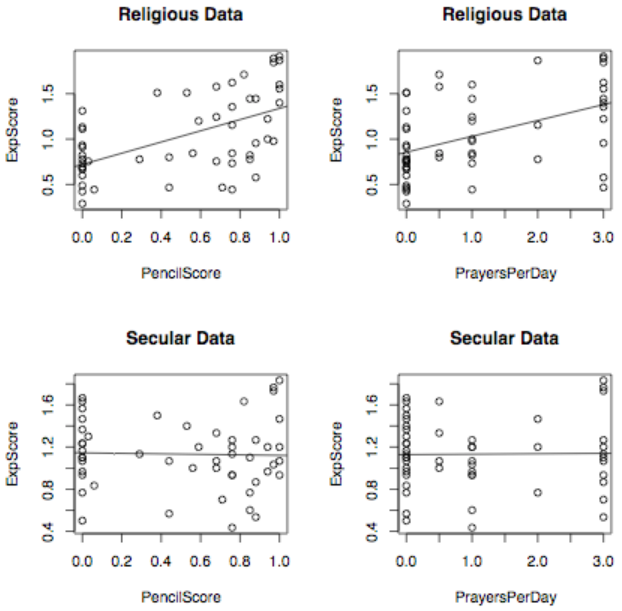


Figure 4: Mean ExpScore per subject as a function of PencilScore and PrayersPerDay (each point represents one subject). Linear models fit to this data suggest that these predictors have a facilitatory effect on the recognition of religious phrases but not of the secular phrases.

Looking first at the subjects' performance on religious phrases, we found that religious subjects obtained higher scores than secular subjects (see Figure 2). This effect was revealed as significant in a 2 x 2 x 3 (Sex x Religion x Frequency) mixed design ANOVA. A by-subject ANOVA revealed main effects for Religion (0.78 vs. 1.2 score for secular vs. religious, $F(1,47)=15.5$, $p<0.001$), Frequency (mean daily score: 1.17, mean weekly score: 0.91, mean annual score: 1.04, $F(2,96)=33.7$, $p<0.0001$), and Sex (0.95 vs. 1.17 score for female vs. male, $F(1,47)=4.5$, $p=0.04$). Interaction effects were found for Religion and Sex ($F(1,47)=4.4$, $p=0.04$), and Religion and Frequency ($F(2,96)=5.5$, $p<0.01$). A by-item ANOVA revealed main effects for Religion ($F(1,42)=124.3$, $p<0.0001$) and Sex ($F(1,42)=46.2$, $p<0.0001$) and a marginal effect for Frequency ($F(2,42)=3.0$, $p=0.06$). An interaction was found for Religion and Frequency ($F(2,42)=3.6$, $p=0.04$). Thus, Religion, Sex and the interaction between Religion and Frequency were significant both by-subject and by-item. Frequency was highly significant by-subject and marginal by-item.

A 2 x 3 ANCOVA with Frequency and Sex as factors and PencilScore as a covariate showed that subjects with high scores in the paper-and-pencil test tended to get high scores in the perceptual identification task (see Figure 4). A by-subject ANCOVA revealed main effects for PencilScore ($F(1,47)=23.8$, $p<0.0001$) and Frequency ($F(2,94)=32.9$, $p<0.0001$). Interactions were found for PencilScore and Sex ($F(1,47)=4.4$, $p=0.04$) and PencilScore and Frequency ($F(2,94)=4.9$, $p<0.01$). A by-item ANCOVA revealed main effects for PencilScore ($F(1,42)=161.2$, $p<0.001$) and Sex ($F(1,42)=29.9$, $p<0.0001$), and a marginal effect for Frequency ($F(2,42)=3.0$, $p=0.06$). An interaction between PencilScore and Sex was also found ($F(1,42)=7.4$, $p<0.01$). Thus, PencilScore and the interaction between PencilScore and Sex were significant both by-subject and by-item. Frequency was significant by-subject and marginal by-item.

A similar 2 x 3 ANCOVA with the same factors but PrayersPerDay as a covariate showed that subjects who pray more tended to obtain better scores (see figure 4). A by-subject ANCOVA revealed main effects for PrayersPerDay ($F(1,47)=15.0$, $p<0.001$) and Frequency ($F(2,94)=32.8$, $p<0.0001$), and an interaction between PrayersPerDay and Frequency ($F(2,94)=5.1$, $p<0.01$). A by-item ANCOVA revealed main effects for PrayersPerDay ($F(1,42)=47.5$, $p<0.0001$) and Sex ($F(1,42)=24.0$, $p<0.0001$), the same marginal effect for Frequency, and an interaction between PrayersPerDay and Sex ($F(1,42)=13.8$, $p<0.001$). PrayersPerDay was significant both by-subject and by-item, and Frequency was significant by-subject and marginally significant by-item.

We performed similar ANOVAs on the secular phrases. The 2 x 2 x 3 ANOVA with factors Sex, Religion, and Frequency revealed that common phrases received higher scores than rare phrases (see Figure 6). A by-subject ANOVA revealed a main effect for Frequency (rare phrases: 0.78, common phrases: 1.48, $F(1,48)=470.7$, $p<0.0001$) and an interaction between Religion and Sex ($F(1,47)=5.6$, $p=0.02$). A by-item ANOVA revealed main effects for Frequency ($F(1,28)=25.8$, $p<0.0001$) and Sex ($F(1,28)=14.6$, $p<0.001$) and an interaction between Religion and Sex ($F(1,29)=15.7$, $p<0.001$). Frequency and the interaction between Religion and Sex were significant both by-subject and by-item (see figure 3). The 2 x 3 by-subject ANCOVA with Sex and Frequency as factors and PencilScore as a covariate found a main effect for Frequency ($F(1,47)=473.6$, $p<0.0001$) and an interaction between PencilScore and Sex ($F(1,47)=5.7$, $p=0.02$). The by-item ANCOVA revealed main effects for Frequency ($F(1,28)=25.8$, $p<0.0001$) and Sex ($F(1,28)=14.4$, $p<0.001$) and an interaction between PencilScore and Sex ($F(1,28)=19.13$, $p<0.001$). Frequency and the interaction between PencilScore and Sex were significant both by-subject and by-item. The last 2 x 3 by-subject ANCOVA with Frequency and Sex as factors and PrayersPerDay as a covariate found a main effect for Frequency ($F(1,47)=489.0$, $p<0.0001$) and an interaction between PrayersPerDay and Sex ($F(1,47)=5.7$, $p=0.02$). The by-item ANOVA revealed main effects for Frequency ($F(1,28)=25.8$, $p<0.0001$) and Sex ($F(1,28)=20.8$, $p<0.0001$), and interactions between PrayersPerDay and Sex ($F(1,28)=15.9$, $p<0.001$) and between PrayersPerDay, Frequency and Sex ($F(1,28)=5.1$, $p=0.03$). Frequency and

the interaction between PrayersPerDay and Sex were significant both by-subject and by-item.

As mentioned earlier, showing significance of a predictor by-subject and by-item does not necessarily entail that it would prove significant in a combined random effects analysis. The lme4 package handles crossed random factors such as Subject and Item in the present case; we used it to fit a linear model to our data. In addition to offering a more faithful picture of the significant effects, the lmer function in this package (but not the standard R function for ANOVA, aov) tolerates unbalanced data (in our case, the numbers of subjects in the Religion by Sex groups were different). The lmer function also allows one to specify a distribution other than the normal distribution for the data. Indeed, the ExpScore variable in our experiment was binomial rather than normal, under a transformation that mapped all erroneous and partial identifications scores to 0, and all perfect scores to 1.

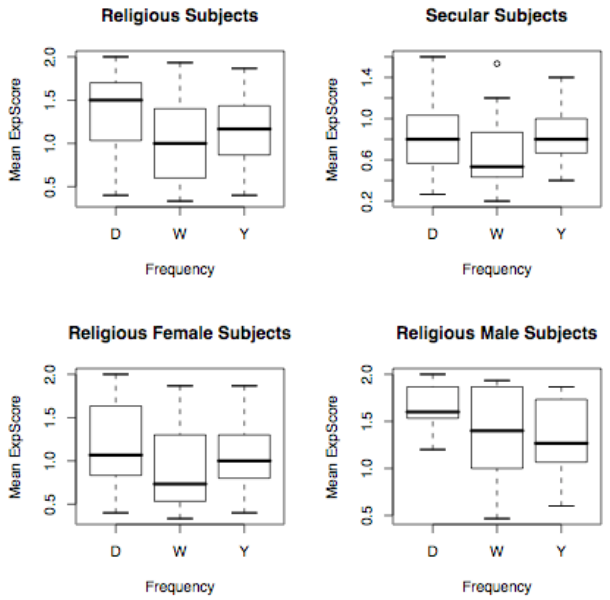


Figure 5: Mean ExpScore per subject across the different levels of frequency for the religious phrases. The graded effect is stronger for religious compared to secular subjects, and for religious men compared to religious women.

We fit a series of binomial logit-link mixed linear models to the ExpScore data for the religious phrases among all of the subjects. The first model used Religion and Frequency as predictors. The Religion predictor was significant, which shows that religious subjects indeed performed better than secular subjects ($z = 5.16$, $p<0.0001$). The second model used PencilScore and Frequency as predictors. The PencilScore predictor was significant, which shows that high scores in the paper-and-pencil test are associated with high scores in the perceptual identification experiment ($z=5.87$, $p<0.0001$). The third model used PrayersPerDay and Frequency as predictors. PrayersPerDay was significant, which confirms that subjects who pray more performed better in the experiment ($z = 4.98$, $p<0.0001$). Sex was dropped from all the above models because its effects did not reach significance when incorporated into the models. Together, the three models show that the more religious the subject is (by whichever of the three measures), the better score received for religious phrases in the experiment.

We next used mixed-model linear regression analysis to test for a graded effect of the frequency of religious phrases among the religious subjects only. Fitting a linear model to the set containing only religious subjects (see Figure 5) showed a main effect for the frequency of the religious phrases ($z=-2.42$, $p=0.015$), and a main effect for sex ($z = 3.11$, $p<0.01$). Because religious men pray more regularly than religious women (mean score for religious males: 1.43, mean score for religious females: 1.03) we limited the analysis further to religious males only and found a stronger graded effect ($z = -3.26$, $p=0.001$). A model fit to the secular subjects data showed no graded effect for the frequency of religious phrases, as expected ($z = -1.14$, $p=0.25$). Adding a quadratic Frequency term to the model for the secular subjects data resulted in marginal effects for Frequency ($z=-1.93$, $p=0.053$) and Frequency squared ($z=1.78$, $p=0.075$).

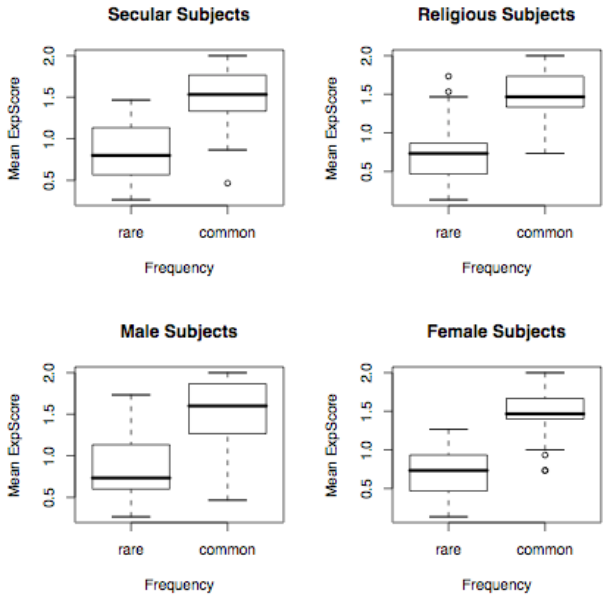


Figure 6: Mean ExpScore per subject across different levels of frequency for the secular phrases. The graded effect is clear regardless of sex or religion.

Models fit to the secular phrases showed a strong effect for the Frequency predictor. A model using Religion, Sex and Frequency as predictors revealed a main effect for Frequency ($z=4.57$, $p<0.0001$) and an interaction between Religion and Sex ($z=2.81$, $p<0.01$). A model using PencilScore, Sex and Frequency as predictors revealed a main effect for Frequency ($z=4.81$, $p<0.0001$) and an interaction between PencilScore and Sex ($z=2.92$, $p<0.01$). A model using PrayersPerDay, Sex and Frequency as predictors revealed again a main effect for Frequency ($z=4.73$, $p<0.0001$) and an interaction between PrayersPerDay and Sex ($z=3.16$, $p<0.01$).

## Discussion

The present study tested the hypothesis that multi-word phrases have increased accessibility as a function of how frequently they are encountered. To quantify prior exposure, we recruited observant religious subjects, who are likely to recite liturgical texts at prescribed regular intervals; secular subjects, who do not pray regularly, served as a control population.

Our results support the graded entrenchment hypothesis. The effects of all three variables that we chose as surrogates for prayer frequency (Religion, PencilScore and PrayersPerDay) on the subjects' recognition performance as measured by ExpScore were significant. We note that while Religion and PrayersPerDay are based on the subject's self-assessment and may therefore not be entirely reliable, PencilScore is an objective "offline" measure of the subject's familiarity with the phrases, just as ExpScore is an objective "online" measure of the same. The observed correlation between these two measures can be explained in terms of a dual effect of entrenchment: on the one hand, it consolidates the memory trace of a phrase (making its eventual retrieval more likely and more reliable); on the other hand, it also makes this memory trace more readily traversed during perception and production (making its real-time recognition more reliable). As expected from this explanation, the correlation between ExpScore and PencilScore is only observed for religious phrases (Figure 4).

Because religious men usually pray more often than religious women, we expected to find an interaction between Religion and Sex. Although the results do point in that direction (see Figure 2), the interaction was not significant in our analysis. We believe that collecting more data might expose this interaction.

We had expected a graded effect within the religious subjects for the frequency of the phrases, and no effect for the secular subjects. However, as Figure 5 shows, the secular subjects did show an effect: a lower score for the weekly phrases compared to the daily and annual phrases. It is possible that the weekly phrases we used were unintentionally more difficult than the daily and annual phrases. The explanation we favor is that non-religious participants are affected by the frequency of phrases, just as are the religious participants, but in an attenuated manner. The annual phrases were more entrenched than weekly phrases within the secular subjects because many secular subjects attend synagogue during the High Holy Days, and because the prayers recited on that occasion carry a high emotional charge.

Within the religious subjects the situation is clear-cut. The model's fit to all of the religious subjects, and especially the fit to the religious men's data, shows sensitivity to the frequency of the phrase. Thus, we may conclude that differential exposure to multi-word phrases is capable of bringing about a graded entrenchment of these phrases — laying down tracks in the subject's mind that facilitate subsequent traversal of familiar phrasal territory.

## Acknowledgments

We thank Giles Hooker for supplying invaluable statistical advice, Ben Sandbank for insightful discussion, and the subjects who participated in this research. JB is a fellow in The Azrieli Fellows Program.

## References

Baayen, R. H. (in press). *Analyzing linguistic data. A practical introduction to statistics using R.* Cambridge: Cambridge university press.

Becker, J. D. (1975). The phrasal lexicon. *Proceedings of the 1975 workshop on theoretical issues in natural language processing.* (pp. 60-63). Morristown, NJ: Association for Computational Linguistics.

Clark, H. H. (1973). The language-as-fixed-effect-fallacy: a critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behaviour,* 12, 335-359.

Eskildsen S. & Cadierno T. (2007). Are recurring multi-word expressions really syntactic freezes? Second Language Acquisition (SLA) from the perspective of Usage-Based Linguistics. *Collocations and Idioms 1. Papers from the First Nordic Conference on Syntactic Freezes.* Joensuu, Finland: Joensuu University Press.

Goldberg, A. (2006). *Constructions at work: the nature of generalizations in language.* Oxford: Oxford University Press.

Harris, C. L. (1998). Psycholinguistic studies of entrenchment. *Conceptual Structures, Language and Discourse*. Berkeley, CA: CSLI.

Caldwell-Harris, C. L. and Morris, A. L. (under review). Fast Pairs: a visual word recognition paradigm for measuring entrenchment, top-down effects, and subjective phenomenology.

Langacker, R. W. (1987). *Foundations of cognitive grammar*. Stanford, CA: Stanford University Press.

Raaijmakers, J. G. W., Schrijnemakers, J. M. C. & Gremmen F. (1999). How to deal with "The language-as-fixed-effect fallacy": common misconceptions and alternative solutions. *Journal of Memory and Language,* 41, 416-426.

Sinclair, J. (1991). *Corpus, concordance and collocations.* Oxford: Oxford University Press.

Theakston, A. L. (2004). The role of entrenchment in constraining children's verb argument structure overgeneralizations: a grammaticality judgment study. *Cognitive Development,* 19, 15-34.

Tomasello, M. (2003). *Constructing a language: a usage-based theory of language acquisition.* Cambridge, MA: Harvard University Press.

Yang, C. (2005). On productivity. In P. Pica, J. Rooryck and J. Van Craenenbroeck (eds.) *Linguistic Variation Yearbook 2005* (pp. 265-302). Amsterdam: John Benjamins Publishing Company.