Cognition 155 (2016) 204–232

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Original Articles

Investigating locality effects and surprisal in written English syntactic choice phenomena

Rajakrishnan Rajkumar [a,*], Marten van Schijndel [b], Michael White [b], William Schuler [b]

[a] Department of Humanities and Social Sciences, IIT Delhi, Hauz Khas, New Delhi 110016, India
[b] Department of Linguistics, The Ohio State University, Oxley Hall, 1712 Neil Ave., Columbus, OH 43210, USA

ARTICLE INFO

Article history:
Received 31 December 2015
Revised 31 May 2016
Accepted 14 June 2016

Keywords:
Language production
Dependency locality
Surprisal
Constituent ordering

ABSTRACT

We investigate the extent to which syntactic choice in written English is influenced by processing considerations as predicted by Gibson's (2000) Dependency Locality Theory (DLT) and Surprisal Theory (Hale, 2001; Levy, 2008). A long line of previous work attests that languages display a tendency for shorter dependencies, and in a previous corpus study, Temperley (2007) provided evidence that this tendency exerts a strong influence on constituent ordering choices. However, Temperley's study included no frequency-based controls, and subsequent work on sentence comprehension with broad-coverage eye-tracking corpora found weak or negative effects of DLT-based measures when frequency effects were statistically controlled for (Demberg & Keller, 2008; van Schijndel, Nguyen, & Schuler 2013; van Schijndel & Schuler, 2013), calling into question the actual impact of dependency locality on syntactic choice phenomena. Going beyond Temperley's work, we show that DLT integration costs are indeed a significant predictor of syntactic choice in written English even in the presence of competing frequency-based and cognitively motivated control factors, including *n*-gram probability and PCFG surprisal as well as embedding depth (Wu, Bachrach, Cardenas, & Schuler, 2010; Yngve, 1960). Our study also shows that the predictions of dependency length and surprisal are only moderately correlated, a finding which mirrors Dember & Keller's (2008) results for sentence comprehension. Further, we demonstrate that the efficacy of dependency length in predicting the corpus choice increases with increasing head-dependent distances. At the same time, we find that the tendency towards dependency locality is not always observed, and with pre-verbal adjuncts in particular, non-locality cases are found more often than not. In contrast, surprisal is effective in these cases, and the embedding depth measures further increase prediction accuracy. We discuss the implications of our findings for theories of language comprehension and production, and conclude with a discussion of questions our work raises for future research.

1. Introduction

A long line of previous research, comprising both spontaneous production experiments and corpus analyses, has studied the production biases involved with constituent ordering. In general, languages are attested to favor producing shorter dependencies, as Liu (2008) demonstrates in a cross-linguistic study involving twenty languages. Fig. 1 shows this trend for English using data from two corpora, the Brown corpus (Francis & Kučera, 1989) and the Wall Street Journal (WSJ) portion of the Penn Treebank (PTB; Marcus, Marcinkiewicz, & Santorini, 1993).

In this paper, we investigate whether this generalization holds true for constructions where speakers have a choice of expressing the same idea using competing word orders, as in the following example (italics added):

(1) a. One day Maeterlinck, coming *with a friend* upon an event which he recognized as the exact pattern of a previous dream, detailed the ensuing occurrences in advance so accurately that his companion was completely mystified. (Brown corpus CF03.10.0)
b. One day Maeterlinck, coming upon an event which he recognized as the exact pattern of a previous dream *with a friend*, detailed the ensuing occurrences in advance so accurately that his companion was completely mystified. (Constructed alternative)

* Corresponding author.
E-mail addresses: raja@iitd.ac.in (R. Rajkumar), vanschm@ling.osu.edu (M. van Schijndel), mwhite@ling.osu.edu (M. White), schuler@ling.osu.edu (W. Schuler).

http://dx.doi.org/10.1016/j.cognition.2016.06.008
0010-0277/

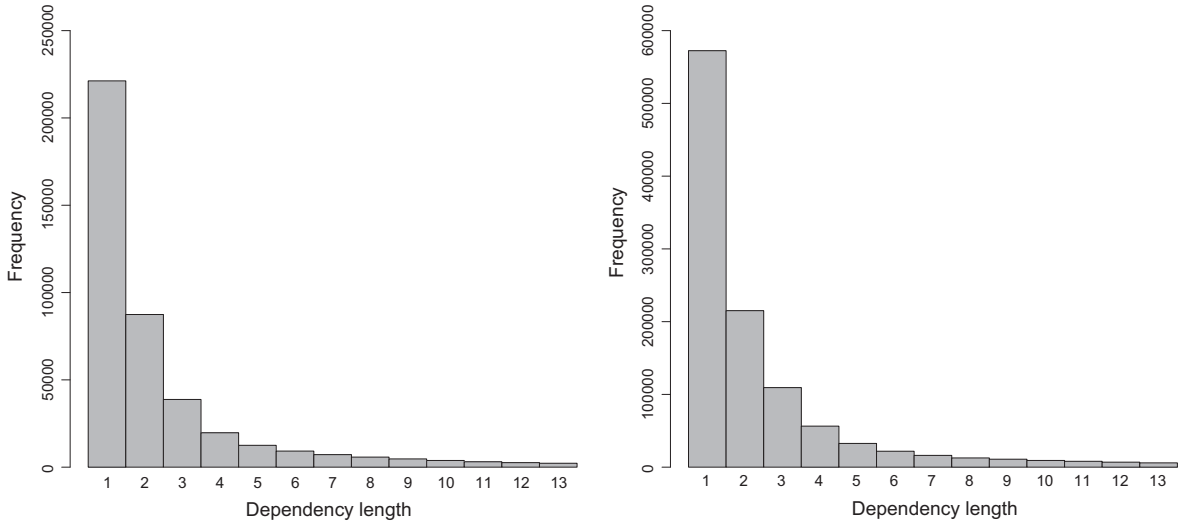

Fig. 1. Dependency length distributions.

Research in the past decade has investigated the hypothesis that one of the factors which influences the structuring of languages is the ease of comprehension and production, in addition to abstract learning biases in language acquisition (Chater & Christiansen, 2010; Hawkins, 2004, 2014). More concretely, do speakers display a preference to produce (1-a) above, since it is easier to produce or comprehend compared to (1-b)? Using a corpus study, Temperley (2007) showed that the tendency to minimize DEPENDENCY LENGTH has a strong influence on constituent ordering choices in written English. In the corpus sentence (1-a), there is a short intervening adjunct *with a friend* between the verb *coming* and the subsequent long constituent starting with *upon*, thus inducing a shorter dependency in comparison to the competing order in the constructed alternative (1-b). Moreover, it is easy to misparse the variant as having *previous dream with a friend* as a constituent, even though this gives rise to a nonsensical interpretation where the dream is a joint activity with the friend.

Dependency length minimization has a long history in the literature dating back to Behaghel's (1932) principle of end weight. In a long line of pioneering work, Hawkins has shown that languages tend to prefer shorter dependencies (Hawkins, 1994, 2000, 2001, 2004, 2014). In the context of syntactic choice phenomena like heavy NP shift (Arnold, Wasow, Losongco, & Ginstrom, 2000; Wasow, 2002), dative alternation (Bresnan, Cueni, Nikitina, & Baayen, 2007), verb-particle shifts (Hawkins, 2011; Lohse, Hawkins, & Wasow, 2004) and topicalization and left-dislocation (Snider & Zaenen, 2006), many other works also corroborate the tendency of languages to minimize dependency length. There is cross-lingual evidence that word order patterns in SOV languages conform to dependency locality (Hawkins, 1994, 2004). The definition of Early Immediate Constituents (EIC) in Hawkins (1994) predicts that for verb-final languages, long constituents tend to precede short ones in the preverbal position. He validates his prediction using Japanese data, and subsequent research builds on EIC predictions in language production studies in Japanese (Yamashita & Chang, 2001) and Korean (Choi, 2007). There is also parallel evidence from optional function words, which are likely to be omitted to shorten dependencies (Hawkins, 2001, 2003; Jaeger, 2006, 2010, 2011).

Temperley's (2007) corpus study uses a variant of Gibson's Dependency Locality Theory (DLT; Gibson, 1998, 2000), a resource-limitation theory of human sentence comprehension, to account for a wide variety of syntactic choice constructions in two written English corpora. Crucially, Temperley's corpus study does not control for other possible explanations of syntactic choice aside from DLT; in particular, it includes no frequency-based controls. Explaining syntactic choice data in terms of a single factor (viz. length or dependency minimization) has also been criticized as being reductive (Bresnan et al., 2007; Snider & Zaenen, 2006; Wasow, 2002). Some corpus studies on specific constructions either hold frequency constant or control for it with lexical counts, as in the case of studies on Heavy NP shift (Arnold, Wasow, Asudeh, & Alrenga, 2004; Arnold et al., 2000), dative alternation (Bresnan et al., 2007), object relative clauses (Jaeger, 2006), complement clauses (Jaeger, 2010) and subject relative clauses (Jaeger, 2011). These studies provide preliminary evidence that dependency length is a significant predictor of ordering choices even when frequency-based controls are considered.

However, in sentence comprehension, although dependency length has been shown to correlate with reading times on constructed stimuli (Levy, Fedorenko, & Gibson, 2013; Warren & Gibson, 2002), it has been difficult to replicate this effect in broad-coverage naturalistic data as strong statistical frequency controls reduce or reverse the effect of dependency length (Demberg & Keller, 2008; Shain, van Schijndel, Gibson, & Schuler, 2016; van Schijndel & Schuler, 2013; van Schijndel, Nguyen, & Schuler, 2013).[1] Even when previous production studies have used explicit frequency controls, they have only used frequency information about individual lexical items and the frames those items occur in, which may not be sufficient. For example, van Schijndel, Schuler, and Culicover (2014) demonstrated that the structural bias statistics captured by latent-variable PCFGs are at least as strong a frequency confound in comprehension as the information captured by lexical counts and subcategorization frame frequencies. Importantly, the structural biases they examine stem from underlying syntactic configurations which may not be readily apparent when counting the number of times a given lexical item occurs in a certain frame (e.g., the probability of a gap being passed into a left branch compared with a right branch at each point in the syntax tree is

[1] As one of the reviewers pointed out, frequency can be considered as an interesting factor in its own right. Please refer to Table 6.1 of MacDonald (1999) which points to many works which consider frequency in production and comprehension research.