Session 4 - Investigating language production at the discourse level

---

We started the session by looking at a generic model of language production. There are many models of language production in the literature, each of these incorporating different insights about possible constraints and interactions between different levels of linguistic organization as well as between language and other cognitve domains. Our point of departure is therefore a stripped down version of what connects different psycholinguistic models of production. There are three core stages to any characterization of language production:

- **Conceptualization** - The process of delineating the message to be conveyed
- **Utterance formulation** - The process of choosing how the message will be instantiated linguistically
  - **Lexicalization** - The process of selecting the lexical items/ constructions to be articulated
  - **Syntactic structuring** - The process of organizing the selected lexical items/ constructions in a meaningful and contingent order
- **Articulation** - The process of executing the motor movements involved in the realization of the linguistic signal (be it acoustic or visual[-spatial])

Signal reduction

The languages of the world offer their users a vast and flexible array of options to linguistically encode their intended messages. In particular, language users have the flexibility to modulate the amount and quality of the linguistic signals they produce, such that they can communicate near meaning-equivalent messages with remarkble variation in the linguistic forms they employ. In this session, we explored the phenomenon of probabilistic reduction, which is the tendency that speakers have to use shorter linguistic forms and more reduced linguistic signals for contextually predictable message parts. We use this exploration to gain some insights about the architecture underlying the language production system. We started by discussing what exactly gets reduced when one talks about linguistic reduction, and for that we distinguished between three constructs which directly relate to our initial model of production: **message components**, which are the parts, broadly construed, of the message the speaker wishes to convey; **linguistic forms**, which are instances of different linguistic categories (e.g. phonemes, words, phrases) that despite not being directly observable in the world are deemed to underlie the signals produced by language users; and lastly **linguistic signals**, which are the observable product of the production process, or in other words, the acoustic and visual signals articulated by motor movements of the humann body. Having clarified what can get reduced in the production of linguistic signals, we then turned our attention to empirically attested instances of probabilistic reduction which relate to pragmatics and discourse. More concretely, we looked at the reduction and omission of referring expressions as well as at reduction beyond the clause level.

Empirical evidence of probabilistic reduction

A long-debated question in language production research is how do speakers decide between near-meaning equivalent forms. For instance, when referring to a work colleague named John, how does one choose between using a pronoun (“he”), a name (“John”), or a full lexical noun phrase (“a colleague of mine”)? How accessible the referent is in a particular context is undoubtedly a crucial deciding factor. However, how can one account for context accessibility in a systematic way? One way of doing that is by looking at predictability in context. Research has shown that the previous mention of a referent impacts how likely it is to be referred again

but also how it might be referred to in subsequent mentions. Specifically, previous mentions increase the likelihood of future mentions and the likelihood of the referent being referred to via a reduced form (Bard et al., 2000). While the likelihood of a new mention decreases with distance to the original mention, longer distances to the original mention also correlate with a lower preferece for pronouns (Arnold, 1998; Arnold, Bennetto, & Diehl, 2009). Similarly, while less predictble preceding context has been linked to the selection of longer linguistic forms (Tily & Piantadosi, 2009), higher contextual predictability correlates with a preference for shorter forms (Mahowald et al., 2013).

When it comes to reduction that extends beyond the level of the clause, it has been shown that coherence markers are more likely to be omitted in cases where there is a contextual cue to a discourse relation that is congruent with the semantics of a given marker (e.g. the marker *instead* is likely to be omitted if there is a cue to an unexpected discourse relation, Asr & Demberg, 2015). The preference for encoding a message in a single vs. multiple clauses is also known to be dependent on the predictability of certain words, such that referents with less predictable labels might be referred to in a split construction as opposed to a single clause (e.g. *Move the triangle to ...* vs. *Take the triangle. Now move it to ...*, Gallo et al., 2008).

The empirical findings reported in the literature strongly suggests that there is an inverse correlation between predictability and linguistic signals, such that less predictable contexts correlate with the usage of longer linguistic forms and overall longer linguistic signals. Still, several open questions remain regarding the role of predictability and reduction in the production of language. For instance, what is the precise nature of the relation between production planning and the realization of the linguistic signal? Or how does reduction operate at different levels of linguistic encoding? More work is needed in order to understand what sort of cues might affect probabilistic reduction, as previous research has focused primarily on predictability as indexed by local lexical cues. A central topic of research should be how multiple cues to the same target might be integrated at different stages of the production process.

Theoretical accounts of probabilistic reduction

The empirical phenomena described above have been accounted for in various ways, most accounts of reduction processes focusing on one of three dimensions: either the ease with which linguistic signals are produced, the constraints imposed on production by communicatve needs, or the role of long-term, offline changes in the representations that get produced. We focused our attention on online constraints on language production, more specifically on accounts that foreground the role of communication and communicative goals. As we noted, despite certain theories being characterized either in terms of ease of production or in terms of communicative goals, in reality all production accounts lie on a continuum of perspectives, where what varies is the focus given to these competing pressures. For instance, communicative accounts do acknowledge that peope have to deal with limited cognitve resources as they try and achieve their communicative goals. Much the same way, production ease accounts do not discard the possibility that speakers might design their utterances in a listener-specific way, however they maintain that this is resource-demanding.

[**Communicative accounts**] Communicative accounts posit that preferences in signal reduction and enhancement are affected by speakers' communicative goals. This usually assumes a bias for robust signal transmission, where transmission is taken in a very broad sense that also includes the transmission of non-literal and social meaning. In this sense, the speaker's ultimate goal is to cause a change in their interlocutor's state of mind. According to these accounts, understanding production preferences involves understanding and accounting for the goals of language use.

Even within communicative accounts, attention is given to the possible trade-offs between ease of production and speakers' communicative goals. In this sense, production preferences are constrained by constant competition between a bias for robust message transmission, as seen above, and a bias for production ease or effort minimization. While the bias for robust message transmission results in a tedency to conserve the quality of the signal, the bias for production ease results in a tendency to produce shorter and less articulated signals. Phenomena of probabilistic reduction can be seen as the result of the fine-tuning between these two biases, where contextual predictability increases the a priori accuracy and speed of message transmission,

ultimately affecting messages as they are being produced and thus before they are processed and interpreted by an interlocutor. Ultimately, reduction should be understood as the default in predictable contexts, as opposed to non-reduction, as predictable messages need not in principle be made more robust via enhancement of the linguistic signal.

Although the insights behind communicative accounts suggest some degree of audience design in the process of language production, such general view construed in terms communicative effiency is qualitatively different from a strong audience design hypothesis, which has more specific assumptions. Indeed, a strong audience design account goes beyond assuming that speakers adapt their productions so as to increase the probability of successful communication, assuming moreover that speakers integrate listener-specific information during encoding, and that they do so immediately during production. When it comes to integrating listener-specific information, what that means is that speakes take into account knowlege about their interlocutor's perspective. For instance, in the case of lexical encoding, they might vary how they select their referring expressions, perhaps referring to an enity in a way that is in line with their intelocutor's perspective and knowledge state. Audience design can thus be regarded as a feature of language production insofar as speakers need to learn how to best communicate in a given situation. In novel situations, as is the case in most psycholinguistic experiments, speakers are confroted with the task of inferring which variant is more likely to help them achieve their communicative goals. If more reduced variants result in communication failure, speakers are able to adapt subsequent productions so as to incorporate less reduced variants.

Open questions in this line of research have to do with the conditions under which speakers engage in audience design, especially in situations that more closely resemble those where natural communicative behavior occurs. It remains to be understood whether speakers are best modeled as rational or boundedly rational agents, though production research would benefit from modeling the utility of audience design, such that the process of producing language might be thought of in terms of the weighing of costs against expected benefits. All in all, the case of probablistic reduction is informative of the processes constraining language production, particularly regarding the way speakers entertain alternative means of expressing near meaning-equivalent messages. It has shown that speakers are sensitive to contextual predictability, and that both production ease and communicative goals mediate language production.