Information Density and Linguistic Encoding (IDEAL)

Matthew W. Crocker¹ · Vera Demberg¹ · Elke Teich²

Published online: 18 September 2015

© The Author(s) 2015. This article is published with open access at Springerlink.com

Abstract We introduce IDEAL (*Information Density and Linguistic Encoding*), a collaborative research center that investigates the hypothesis that language use may be driven by the optimal use of the communication channel. From the point of view of linguistics, our approach promises to shed light on selected aspects of language variation that are hitherto not sufficiently explained. Applications of our research can be envisaged in various areas of natural language processing and AI, including machine translation, text generation, speech synthesis and multimodal interfaces.

Keywords Language variation · Information theory · Information density · Surprise

1 Introduction

Language exhibits a broad spectrum of variation at all linguistic levels. While there is a substantial body of research on the systematic description of linguistic variation accounting for factors such as region (dialect), social background (sociolect) and situational context (register), there are few

attempts at finding a common explanation: Why do we usually have more than one option of encoding a given message? And how do we choose among the options available to us? These are the questions we address in IDEAL.

For illustration consider the following examples.

- (1) (a) *My boss confirmed that he is absolutely crazy.*
(b) *My boss confirmed he is absolutely crazy.*
- (2) (a) *Where should I put this stuff?*
(b) *Where to put this stuff?*
- (3) (a) *If this method of control were to be used, trains would operate more safely.*
(b) *The use of this control method leads to safer train operation.*

What these pairs of utterances have in common is that they are alternative encodings of the same message where the (b) versions are shorter, more reduced encodings than the (a) versions, which are longer and more expanded. We can observe this pattern of variation across languages and at all levels of the linguistic system—from the phonetic (e.g. shorter vs. longer duration of syllables) and lexical levels (e.g. orthographically shorter vs. longer words) to the syntactic (e.g. presence vs. omission of optional elements) and discourse levels (e.g. use of pronouns vs. full referring expressions).

Our overarching research question is what governs the choice between such encoding options. Specifically, we begin with the hypothesis that such encoding choices serve to optimize *density*—the amount of information conveyed per unit of time. Traditionally, linguistics has associated the informational content of a sentence or discourse with its semantics: The inherent meaning of words and constituents are combined compositionally to determine a sentential or discourse message. Psycholinguistics and computational linguistics, however, have increasingly turned to

✉ Elke Teich
e.teich@mx.uni-saarland.de

Matthew W. Crocker
crocker@coli.uni-sb.de

Vera Demberg
vera@coli.uni-saarland.de

¹ Allgemeine Linguistik, Universität des Saarlandes,
Campus C7, 66123 Saarbrücken, Germany

² Angewandte Sprachwissenschaft, Universität des Saarlandes,
Campus C2 2, 66123 Saarbrücken, Germany

Information Theory as a mathematical framework for objectively quantifying the information conveyed by a linguistic unit (e.g. phoneme or word) as a function of its predictability in context [13], often termed *surprisal*.

IDEAL investigates the hypothesis that processing complexity is indexed by surprisal across linguistic levels and further that linguistic variation may be characterized by the optimal distribution of information across the linguistic signal. We provide a more formal definition of surprisal in Sect. 2, along with a summary of supporting evidence for its linguistic relevance, and then illustrate several different research areas in IDEAL in Sect. 3. We conclude with a brief summary and outlook (Sect. 4).

2 Information Density

Given the varying constraints that preceding context exerts upon what linguistic $unit_i$ (e.g. a phoneme, word, or sentence) may follow, Information Theory [13] defines the amount of information that is conveyed by that unit in terms of *bits*—resulting in a measure commonly known as *surprisal*—using the following formula:

$$Surprisal(unit_i) = \log_2 \frac{1}{P(unit_i|Context)} \quad (1)$$

Two fundamental properties of this characterisation are (1) that linguistic events with low probability convey more information than those with high probability, and (2) the information conveyed by a particular linguistic $unit_i$ —be it a phoneme, word, or utterance—is not determined solely by the unit itself, but crucially by the context in which it occurs. Stated simply, surprisal captures the intuition that linguistic expressions that are highly predictable, in a given context, convey less information than those which are surprising.

At the word level, Hale posits that the cognitive effort associated with comprehending the next word, w_i , of a sentence will be proportional to its surprisal (Eq. 1) [5, 9]. This claim is consistent with a wealth of experimental evidence demonstrating that behavioural (e.g. reading time) and neurophysiological (e.g. event-related brain potentials, ERPs) measures of processing effort are highly correlated with a word's predictability [3, 8, 12, 14].

Assuming there is an upper bound on the cognitive resources that listeners can exploit for decoding the language they encounter, a human communication system that strives for optimal efficiency should encode messages in a manner that distributes information as *uniformly* as possible, over time. More specifically, the Uniform Information Density (UID) hypothesis postulates that encoding mechanisms will seek to avoid peaks and troughs in surprisal—avoiding either overloading the listener or being

Fig. 1 Language use strives for good channel use

uninformative—thus optimizing the transmission of information from speaker to hearer. Figure 1 illustrates the surprisal profile for the linguistic units for two possible encodings. The UID hypothesis asserts that production processes will prefer encodings which distribute surprisal more uniformly across the signal—as in the righthand graph—so as to avoid peaks that may exceed the comprehenders processing, or *channel*, capacity.

Evidence for UID comes from the observation that various syntactic reduction phenomena—such as *that*-complementiser and *that*-relativiser omission, as well as auxiliary contraction—can be explained by a preference for uniformity in surprisal [6, 7]. Other relevant support comes from the observation that speakers take more time to pronounce words when they occur in less predictable contexts [1]. While at the text level, Genzel and Charniak [4] find evidence that information density of sentences can be viewed as uniform when taking context into account, in contrast with the apparent increase in density when context is not considered.

Particularly compelling is recent research demonstrating that the lexica of many human languages have adapted so as to encode words that are more predictable (on average) using shorter forms than less predictable words [11]. The consequence of such a lexicon is precisely to increase uniformity by using longer forms for words that typically convey more information, thus distributing the information over time. Indeed, evidence also suggests that people's decision to say *math* rather than *mathematics* is driven at least partly by the increased predictability of the word in a particular context—that is people use the short form, when the word is less surprising and conveys fewer bits of information [10].

3 Research Programme

Building on the findings outlined above, IDEAL investigates the extent to which surprisal offers a pervasive explanation of language behavior across levels of linguistic representation. To this end, we examine both the mechanisms of encoding and determinants of surprisal in detail by (a) identifying which aspects of linguistic and non-linguistic context are relevant for determining levels of

surprisal and density, and (b) examining the diverse means languages make available for variation in linguistic encoding and thus modulation of surprisal: cross-linguistically, diachronically, and in different genres and registers. Measures of surprisal and processing difficulty can then feed back into models of human linguistic behavior as well as various kinds of computational applications. For instance, in human computer interaction (HCI), they open a path to adaptive technology that can make effective use of the variation available in language, to adapt utterances in a situated setting to a user and the environment.

3.1 Research Areas

The projects in IDEAL are distributed across three research areas: “Situational Context and World Knowledge” (Area A), “Discourse and Register” (Area B) and “Variation in Linguistic Encoding” (Area C). We will below give a few examples of the research carried out in each area.

Projects in Area A explore the extension of the notion of context beyond the strictly linguistic context, to relevant aspects of the current situation (e.g., objects present in the scene, interlocutors’ gestures or gaze) which help us explain linguistic variation observed in situated communication. A particular focus concerns of the role of event knowledge in linguistic choice during (human) production. Here, we test experimentally how exactly human knowledge of typical complex but routine events—like going to a restaurant—is structured, and how humans integrate such *script* knowledge with the information contained in utterances that implicitly rely on it. Challenges addressed in these projects involve the acquisition of event knowledge at a large scale from texts and via human annotation, and the mapping between representations of script knowledge and actual textual references to these events, see for example Fig. 2.

Projects in Area B focus on surprisal in discourse context and different registers and text types. One of the projects, for example, looks at the hypothesis of linguistic densification in the evolution of scientific writing in

Fig. 2 The challenge lies in automatically aligning knowledge about event sequences with natural texts

English (mid 17th century to present), starting from the assumption that shared expertise of the author and their audience affects language use and, over a longer period, drives language change and the evolution of domain-specific language (register) [15]. As scientific activity in a given field develops and becomes more specialized, particular meanings become more predictable (within that scientific field). UID then predicts the emergence of denser encodings for these predictable meanings, which would optimize efficiency in communication.

Projects in Area C focus on testing the effect of linguistic predictability on the expansion or compression of linguistic items in encoding. Such effects can be observed in acoustic-phonetic realization in speech (e.g. shorter vs. longer durations of syllables in speech production) and are investigated in two projects in this area. Another project studies intercomprehension across languages of one family (here: Slavic languages), starting from the assumption that there is a correlation between (the degree of) language relatedness, intercomprehension and surprisal.

3.2 Methods for Measuring Surprisal

In order to address questions about the role of information density in language use and language evolution, it is necessary to accurately quantify the amount of information carried by a linguistic item. Two kinds of methods of quantifying surprisal are employed across the three research areas. Surprisal can be estimated experimentally from human subjects by asking for completions, and observing processing difficulty through eye-tracking and event-related potentials in EEG; and it can be estimated on the basis of probabilities obtained from large corpora representative of different domains. With regard to the latter, the development of language modeling approaches with more sophisticated and extended notions of linguistic context, is a particular research focus. Finally, several projects are also concerned with developing computational models which represent aspects of language structure which is not observable in surface forms, modelling for example the surprisal of a syntactic structure, the thematic fit of words in specific roles when occurring in long-distance dependencies, or the information conveyed by discourse connectives and other linguistic cues about discourse relations.

3.3 Applications in HCI and NLP

The notion of information density and its effect on linguistic encoding also has the potential to improve Natural Language Processing (NLP) applications, such as making effective use of the ability to represent alternative linguistic

encodings of a particular meaning: if we can map high-density and low-density encodings of the same information onto one another, this should also help to improve information retrieval, especially for complex multi-argument events.

Cognitive models of processing difficulty can also be used to inform natural language generation in dialog systems, such that the dialog system can choose the optimal utterance for a given user (e.g., layperson vs. expert) in a situation. For instance, we investigate how to optimally manage the cognitive load induced by a language comprehension task in combination with a driving task (in an automotive simulator) for different user groups (younger vs. older adults). Experimental findings will contribute to the development of a language generation model that adapts linguistic encodings appropriately based on both the immediate setting and cognitive capacity of the listener, see also Demberg et al., this issue.

4 Envoi

The overarching goal of IDEAL is to establish the extent to which variation in language and language use—from phonemes to discourse—can be explained by the pressure to distribute information evenly across the communicative channel. The constituent projects thus contribute towards the development of comprehensive model of language use that unites traditionally different perspectives—from the cognitive and computational to the social and historical.

On the conceptual level, we assume that language is *designed for communication* and that language users are rational in the sense that they want communication to work. Hence, they will adapt their linguistic encodings in the service of successful communication. The notion of information density gives us a basis for investigating language use that is in accordance with these assumptions.

On the methodological level, using information theory as a common formal basis opens up a rich repertoire of computational methods commonly used in computational linguistics and in parts of psycholinguistics to other areas of linguistic investigation, including (cross-linguistic) language variation, language acquisition and language evolution. Also, since information theory is agnostic regarding any particular linguistic theory, different theoretical perspectives can be incorporated in a unified model of language use under this view.

In the present project phase, we focus on selected aspects of language variation and developing a solid repertoire of methods of measuring ID/surprisal, building computational models, and determining the scope of surprisal based accounts. In future phases, we would like to

refine our models and expand their application to a broader spectrum of languages and other areas of linguistic investigation, such as language acquisition and learning, language typology and language evolution.

In a wider perspective, we believe that information theory can help describe and explain (aspects of) human experience more generally, notably its evolutionary, cognitive and social aspects. As language is a major window into human behavior, we would thus hope that the research carried out in IDEAL will provide contributions to the larger endeavor of modeling and explaining human experience.

Acknowledgments IDEAL is funded by the *Deutsche Forschungsgemeinschaft* (DFG) under grant SFB (*Sonderforschungsbereich*) 1102 (www.sfb1102.uni-saarland.de). Support by Cluster of Excellence *Multimodal Computing and Interaction* (MMCI) is also gratefully acknowledged.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. Aylett M, Turk A (2004) The smooth signal redundancy hypothesis: a functional explanation for relationships between redundancy, prosodic prominence and duration in spontaneous speech. *Lang Speech* 47(1):31–56
2. Demberg V, Hoffmann J, Howcroft D, Klakow D, Torralba A (2015) Search challenges in natural language generation with complex optimization objectives. *Künstliche Intelligenz* (in this issue)
3. Demberg V, Keller F (2008) Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* 109:193–210
4. Genzel D, Charniak E (2002) Entropy rate constancy in text. In: *Proceedings of the 40th meeting of the Association for Computational Linguistics, ACL '02*, pp 199–206
5. Hale J (2001) A probabilistic earley parser as a psycholinguistic model. *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies, NAACL '01*, Association for Computational Linguistics, Stroudsburg, PA, USA, pp 1–8
6. Jaeger TF (2010) Redundancy and reduction: speakers manage syntactic information density. *Cogn Psychol* 61:23–62
7. Jaeger TF, Tily H (2011) On language utility: processing complexity and communicative efficiency. *Wiley Interdiscip Rev Cogn Sci* 2(3):323–335
8. Kutas M, DeLong KA, Smith NJ (2011) A look around at what lies ahead: prediction and predictability in language processing. In: M. Bar (ed) *Predictions in the Brain: using our past to generate a future*. Oxford University Press, UK, pp 190–207
9. Levy R (2008) Expectation-based syntactic comprehension. *Cognition* 106(3):1126–1177
10. Mahowald K, Fedorenko E, Piantadosi S, Gibson E (2013) Info/information theory: speakers choose shorter words in predictive contexts. *Cognition* 126:313–318

11. Piantadosi S, Tily H, Gibson E (2011) Word lengths are optimized for efficient communication. *Proc Natl Acad Sci* 108(9):3526
12. Rayner K, Well A (1996) Effects of contextual constraint on eye movements in reading: a further examination. *Psychon Bull Rev* 3:504–509
13. Shannon CE (1948) A mathematical theory of communication. *Bell Syst Tech J* 27(379–423):623–656
14. Smith NJ, Levy R (2013) The effect of word predictability on reading time is logarithmic. *Cognition* 128(3):302–319
15. Teich E, Degaetano-Ortlieb S, Fankhauser P, Kermes H, Lapschynova-Koltunski E (2015) The linguistic construal of disciplinary: a data mining approach using register features. *J Assoc Info Sci Technol JASIST*

Matthew W. Crocker is professor of Psycholinguistics at Saarland University, Deputy Director of the DFG-funded Collaborative Research Center *Information Density and Linguistic Encoding* (SFB 1102), and PI in the Cluster of Excellence *Multimodal Computing and Interaction* (MMCI) and the Marie Curie ITN *LanPercept*. His research combines the behavioural and neurophysiological investigation of real-time language processing, with

the development of computational cognitive models.

Vera Demberg is an independent research group leader at the Cluster of Excellence MMCI at Saarland University, Saarbrücken. She is a PI in the DFG-funded Collaborative Research Center *Information Density and Linguistic Encoding* (SFB 1102). Vera Demberg is interested in computational models of human language processing and their use in natural language processing systems, as well as experimental psycholinguistics.

Elke Teich is professor of English Linguistics and Translation at Saarland University and Director of the DFG-funded Collaborative Research Center *Information Density and Linguistic Encoding* (SFB 1102). She is also currently PI in the Cluster of Excellence *Multimodal Computing and Interaction* (MMCI) and the BmBF-funded CLARIN-D (*Common Language Resources and Technology Infrastructure*) project. Elke Teich's research interests

range from linguistic variation with a focus on scientific language to translation and digital humanities.