Title for display, which can be longer or even contain linebreaks

Author A[a,b] and Author B[a,c]

author.a@A.edu, author.b@b.edu

[a] Institution A
[b] Institution B
[c] Institution C

ARTICLE INFO

Peer-Reviewed By:
Article DOI:
Github repo:
Journal ISSN:

ABSTRACT

Abstract here.

The following content is taken from Chang and DeDeo, "Divergence and the Complexity ofDifference in Text and Culture."

Some figures and citations

Fig. 1 summarizes our discussion up to this point. We began by distinguishing difference and distance, and pointed out the inherent limitation of distance as embodied experience. Below we shall see divergence in action. It will turn out that divergence, and the underlying epistemic notions it tracks, violates the axioms of distance—"symmetry" and "no shortcut" in particular—in crucial and useful ways.

difference

operationalization

operationalization

distance

divergence

embodied experience of space

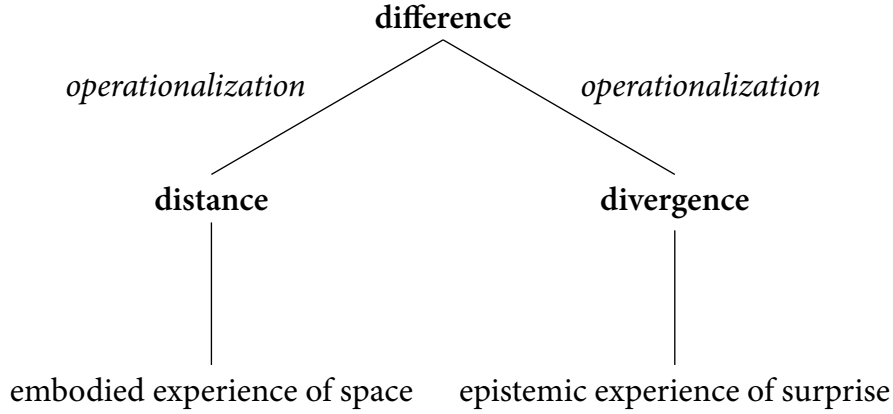epistemic experience of surprise

Figure 1: Two paths to operationalizing difference: a cognitive approach made possible by the information-theoretic notion of divergence provides a different account of how things differ than the more familiar spatial metaphor of distance.

Given an interpretive frame, we can often consider one text as being broader, or ranging over more themes, than another. One can then speak of the asymmetrical relationship of one text being contained, or enclosed, by another. Consider, for example, the case of diction. At any particular time, a culture will have a range of registers available for a text to use.[1] One text may have an elevated diction, another low, and a third might mix the two. A recent study by Stefania Degaetano-Ortlieb and Elke Teich shows how scientific English became broader, or gradually came to *enclose*, general English over the second half of the nineteenth century.[2] There are certainly words that are common in general English that are less common in scientific English (for example, the word "wife"). However, there are (roughly speaking) far more words in scientific English that are rarely, if ever, used in general English (such as "oxide"). This imbalance is measured by the Kullback–Leibler asymmetry.

Some tables

To drive this intuition, consider two major words of Russian literature, *Anna Karenina* and *War and Peace*, by Leo Tolstoy. In a very general sense (and the only one needed to understand this example), both of these novels deal with themes of friendship, suffering, war, and love, but in different propor-

tions; while *Anna Karenina* is most famously centered around a love story, *War and Peace* follows a variety of characters as they navigate a broader array of life events.

We can represent, in this toy example, the two novels as probability distributions over these themes. Explicitly, the distributions p_A (*Anna Karenina*) and p_B (*War and Peace*) quantify the themes by reference to the relative proportions of words associated with the core concepts in order, as in table 1:[3]

$$
\begin{aligned}
p_A &= \{0.14, 0.21, 0.08, 0.57\}, \\
p_B &= \{0.15, 0.15, 0.37, 0.36\}.
\end{aligned}
$$

Intuitively, at this level of analysis, *War and Peace* covers a broader range of themes than *Anna Karenina*; the latter, for example, distributes only 8% of its coverage to war, compared to 37% in *War and Peace*.

theme	*Anna Karenina*	*War and Peace*
friendship	14%	15%
suffering	21%	12%
war	8%	37%
love	57%	36%

Table 1: *While* Anna Karenina *is dominated by words associated with love,* War and Peace *has a more uniform distribution over the four themes considered here. In this sense,* War and Peace *"encloses" the themes in* Anna Karenina, *and while the Kullback–Leibler from* Anna Karenina *to* War and Peace *is 0.49 bits, the Kullback–Leibler in the reverse direction, from* War and Peace *to* Anna Karenina, *is only 0.35 bits. In this example, the compositions of the two novels are determined by counting words associated with the four themes in the English-language translations of Constance Garnett and computing the relative odds. The listed probabilities are then conditional probabilities: given that one has encountered one of these themes, what are the probabilities that it was this one or that.*

The Kullback–Leibler Divergence reflects these differences in the asymmetry between the two directions. The KL from *Anna Karenina* to *War and Peace* (i.e., "the surprise on reading *Anna Karenina* and then encountering *War and Peace*, when one searches for these themes"), is 0.49 bits, while the KL from *War and Peace* to *Anna Karenina* is 0.35 bits. The enclosure relationship is captured in the asymmetry, or difference, in surprise.[4] In this interpretative framework, readers who begin with *Anna Karenina* have little exposure to narratives involving war; when they read *War and Peace*, they encounter new

patterns associated with the war theme more than one third of the time. Conversely, readers who begin with *War and Peace* have significant exposure to all four themes; they see less representation of some of them than they would in *Anna Karenina,* but not enough to overcome the near-total absence of war in the latter. The underlying asymmetry of 0.074 bits between the texts, in the direction from *War and Peace* to *Anna Karenina,* captures this relationship.

Notes

[1] Michail Bakhtin, *The Dialogic Imagination: Four Essays*, trans. Michael Holquist and Caryl Emerson, University of Texas Press Slavic Series 1 (Austin, TX: University of Texas Press, 2011).

[2] See Stefania Degaetano-Ortlieb and Elke Teich, "Toward an Optimal Code for Communication: The Case of Scientific English," *Corpus Linguistics and Linguistic Theory* 0, no. 0 (2019), doi:`10.1515/cllt-2018-0088`, where they establish this "subtle effect" of scientific language enclosing general English in passing. They do so precisely in the way we argue for here, by reference the asymmetric nature of KL divergence. In their phrase, which captures the cognitive aspect of the information theoretic quantities in play, the enclosure relationship indicated by the KL asymmetry implies that general English is a worse "model" for scientific English than vice versa. See also Stefania Degaetano-Ortlieb et al., "A diachronic perspective on efficiency in language use: that-complement clause in academic writing across 300 years," in *Proceedings of the 10th International Corpus Linguistics Conference* (Cardiff, Wales, UK, 2019).

[3] For this toy example, friendship words are: friend, friendship, companionship, fellowship, camaraderie; suffering words are: suffering, hurt, ache, pain, agony, miserable, wretched; war words are: war, conflict, warfare, combat, fighting, struggle, armed, military, bloodshed; and love words are: love, lover, fondness, tenderness, warmth, intimacy, attachment, endearment.

[4] Formally, our operationalization of enclosure, $\mathrm{KL}(p_1|p_2) - \mathrm{KL}(p_B|p_A) > 0$, indicates that p_A encloses p_B.