A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

Laying Algebraic-Geometric Foundation for Probabilistic Models of Natural Languages

Akihiro Maeda

Supervisor: Shohei Hidaka

Japan Advanced Institute of Science and Technology
[Information Science]

March 2026

Abstract

Human language exhibits a remarkable form of compositionality: complex expressions are systematically built from simpler parts, and this structure supports strong out-of-distribution generalization. Despite the empirical success of modern neural language models, no existing approach provides a principled probabilistic account of how such compositional structure is represented, computed, or learned. This dissertation addresses this gap by developing a mathematical framework that treats linguistic probability itself as an algebraic and geometric object.

The first contribution is the formulation of a *structural probability model*, grounded in algebraic statistics. Sentences are represented as joint probability tensors whose algebraic constraints correspond to geometric entities. This provides the first unified framework in which the compositional organization of language is expressed as algebraic system and appears as low-dimensional geometric structure in distributional representations.

The second contribution is the introduction of a new invariant unit of probabilistic structure, termed the *Minimum Invariant Constraint (MIC)*. An MIC is defined as the atomic algebraic component of a probability tensor, mathematically characterized by a vanishing 2×2 minor. MICs generalize the notions of independence; they are invariant under reparameterization, robust across corpora, and serve as the irreducible building blocks for complex structural patterns. This framework provides a principled explanation for the local rank-one patterns observed in PMI matrices and co-occurrence statistics.

The third contribution is the development of computational methods for discovering MICs and their compositions in empirical data. Two complementary approaches are introduced: (i) a divide–and–conquer method based on marginalization identities and cumulants, which interprets PMI heuristics as second-order cumulants; and (ii) a harmonic-analysis method using the Walsh–Hadamard transform and geometric algebra, enabling efficient detection of symmetric and low-rank structure in high-order tensors. Proof-of-concept experiments demonstrate that these methods recover invariant components that were previously inaccessible to conventional tensor decompositions.

Together, these contributions establish a new algebraic–geometric foundation for linguistic compositionality. They show that the internal organization of language emerges as invariant algebraic constraints on probability distributions, offering both theoretical insight into the structure of language and practical tools for analyzing modern language models.

Keywords: language model, algebraic statistic, invariant constraint, compositionality, vanishing binomial