Truth: Between Disquotation and Compositionality

The first half of the twentieth century brought a rapid development of formal logic. Philosophers and mathematicians working hand in hand, succeeded in clarifying many important philosophical concepts. On the one hand, scholars managed to provide a rigorous definitions of notions of more syntactical character, such as *formal language*, *provability* and *computability*. On the other, the work was driven mostly by the discovery of paradoxes of essentially semantical origins. Naïve approaches to the formalization of concepts such as *definability*, *meaning* and *set* quickly led to contradictions (in the form of Richard's paradox, Grelling–Nelson's paradox and Russel's paradox, respectively). This was the case also of the proper topic of our project: the notion of *truth*.

Arguably the most basic principles governing the use of the expression "is true" are given by the statements of the form ""A" is true if and only if A.", where A is a sentence. Each such statement is called a *T–biconditional* and their collection: a *disquotational scheme*. This collection as a whole expresses that the role of the notion of truth is to invert the effect of taking a sentence into the quotation marks, hence to *disquote*. As famously demonstrated by Tarski (basing on an earlier work of Gödel) if in the above scheme A is allowed to contain the expression "is true", then the disquotational scheme is *inconsistent*. The proof of this fact proceeds by formalizing the well known (discovered already in antiquity) *liar paradox*: one considers a sentence λ expressing ""λ" is not true" and by applying the T–biconditional for λ arrives at a contradiction. Essentially this is the content of Tarski's *undefinability of truth* theorem.

However, by disallowing the expression "is true" to occur in sentences to which the notion of truth is applied (in particular to λ, as introduced above), one is able to block the reasoning from the liar paradox. In such a situation, Tarski showed that, it is even possible to *define*, for a given language $\mathscr{L}$, the expression "is true in $\mathscr{L}$" in such a way that all T–biconditionals for sentences of $\mathscr{L}$ will follow. Let us stress that this is not entirely trivial, since the disquotational scheme contains infintiely many sentences each of which being formally independent (i.e. neither implying nor being implied) of the others. The method to derive infinitely many T–biconditionals from finitely many clauses consisted in making use of the *compositional principles* for $\mathscr{L}$. They can be thought of instructions showing how the truth of a compound sentence depends on the truth of its constituents. For example, if $\mathscr{L}$ is (a fragment) of the English language, then the compositional clause for the connective "and" reads as follows

For all sentences ϕ and ψ, the sentence "ϕ and ψ" is true if and only if ϕ is true and ψ is true.

Already Tarski observed that finitely many conditions of the above form (for all connectives and quantifiers of the language $\mathscr{L}$) are much more expressive than the disquotational scheme (for the language $\mathscr{L}$). Crucially, they enable us to justify that truth is preserved in formal deductions, which is a highly desirable consequence of the theory of truth. The disquotational scheme alone falls short of satisfying this requirement, which in more modern philosophical literature, formed a serious argument against its completeness as an axiomatization of the notion of truth.

In the project, we investigate the relationship between two above mentioned principles for the notion of truth: the *disquotational scheme* and the *compositional principles*. In particular we seek to understand whether the compositional clauses, despite their unprovability from the disquotational scheme, can in some sense be founded on T–biconditionals. The crucial philosophical motivation for this question is provided by the main thesis of *disquotationalism*. The proponents of this stance claim that, despite their formal weakness, T–biconditionals suffice to capture all the essence of the notion of truth. Secondly, we ask whether compositionality is *necessary* to express the infinite disquotational scheme with the use of only finitely many conditions. Last but not least, we investigate the generalisations of Tarski's undefinability of truth theorem where the definability condition is relaxed to interpretability.

The above questions will be investigated by applying formal methods of *axiomatic theories of truth*. This is a subfield of formal logic, which thus far provided many important insights regarding the formal properties of the notion of truth, for example showing which collections of principles for this notion are consistent. Each of the above mentioned questions admits very natural counterparts in the form of mathematical hypotheses which we will seek to confirm or refute. The formal results will guide us in verifying our philosophical intuitions and formulating new hypotheses. The project expected outcome consists in the solution of certain, interesting on their own, formal problems and, as a consequence, clarifying our intuitions regarding one of the most philosophically relevant notions.