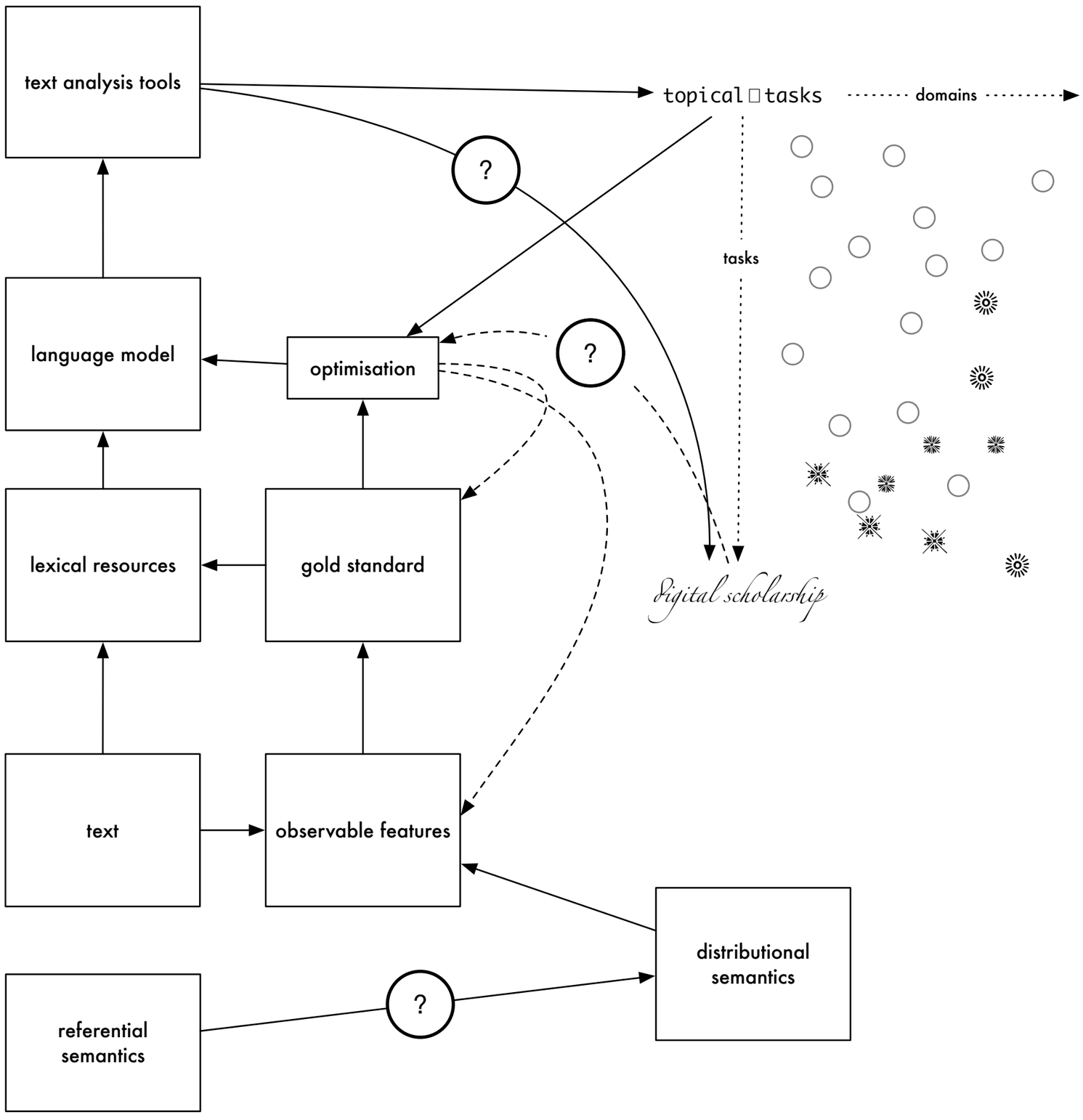
Lexical Gold Standards, Text Analysis Tools, Digital Scholarship

Jussi Karlgren, Gavagai

9 September 2019, Lugano



Referential Semantics

Referentiality covers one of the more important aspects of language use: that of topicality, where language calls up items, concepts, notions of interest to discourse participants. In most computational text analysis tasks, *topicality* has been at the center of attention: what a text is *about* is the primary categorisation criterion.

The general intuition of topical analysis is that many terms in language appear in a tight bursty pattern to indicate that some matter of interest is under treatment, and that other terms appear in a wider distribution, constituting structural material rather than topical. As an example, texts which contain terms *helicopter*, *rotor*, *airfield*, and *pilot* vs texts which contain the terms *cow*, *milk*, *dairy*, and *barn* can with some ease be classified topically from bursty term occurrence alone. Terms such as *see*, *move*, *rotate*, or *yield* are not as useful for this purpose. This gives us quality criteria for text analysis tools related to coverage over bursty terms and filtering mechanisms for widely dispersed terms. Finding topical synonyms (*autogiro*, *chopper*, *whirlybird*) or topically related terms (*airfoil*, *camber*, *translational lift*) and removing structural, perspective-related, or attributive terms are worthwhile efforts to improve results.

Digital Scholarship

New tools, new methods, and new results allow scholars to work with entire collections rather than small selected sets of cultural items. The new methodologies *distant reading*, to complement the traditional close reading and new tools enable scholars to find new types of patterns in their research material. But humanities and the social sciences approach their material in ways which are different from the tasks text analysis methods are built to accommodate. Questions of interest are typically not what a text is about, but e.g.: how authors and schools of thought spread and influence each other, how much or little knowledge of distant cultures there was at some time in some cultural area, how political institutions change over time, how argumentation influences decision making, how public sentiment affects financial indicators, how the well-being of individuals are manifested in their writing, how a scholarly field selects its focus topics, how language change is motivated by local prestige markers, how social change is reflected in literary work, how to determine who has authored a given work, and so forth.

Many of these questions are well served by large scale work on large collection, and many of these questions have been touched upon or addressed directly in recent years in experimental work here at CLEF: they necessitate new evaluation schemes. And those schemes need other resources.