Commentary

Distilling a Materials Synthesis Ontology

Edward Kim [1], Kevin Huang [1], Olga Kononova [2], Gerbrand Ceder [2], and Elsa Olivetti [1]*

[1] Dept. of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA
[2] Dept. of Materials Science and Engineering, University of California, Berkeley, CA, USA
* Correspondence: elsao@mit.edu

When reading a journal article describing the synthesis of a material, the key results are easily found: Was a new morphology achieved? Or has a functional performance barrier been broken? A quick skim, perhaps of the title alone, provides these answers. Yet, extracting the essence of the experimental synthesis *method* is far more difficult. There is no standardized structure for written methods sections, beyond the common practice of using past tense narratives with a passive voice. As we enter the age of accelerated materials screening, rapid first-principles computations, and massive structure-property databases, the rate-limiting step for materials development has become the discovery and validation of synthesis methods [1]. We should take steps to ensure that the community is not writing itself into a dead end.

Despite the ever-expanding volume of published literature, data-driven materials science has been enabled largely by the proliferation of machine-readable, curated datasets. The synthesis of organic molecules, for example, has recently been predicted at human-level accuracy by algorithms trained on millions of historical reactions [2]. On the other hand, data-driven *inorganic* synthesis has yet to see AI-guided results analogous to that of Segler et al. [2], as no comprehensive database of codified inorganic synthesis has been created [1]. The vast majority of inorganic materials syntheses are recorded solely in the methods sections of journal articles, and our ability to harness this knowledge in its entirety is ultimately gated by the writing styles used in the research community.

Efforts to text-mine materials science and chemistry literature have nonetheless made progress [1,3], but accurately codifying entire synthesis routes, using only the original written text as input, is still an unsolved problem. Machine-learning-guided inorganic synthesis, using algorithms trained on text-based data, has so far only been realized with the aid of manual data extraction [4]. To effectively search the vast space of materials synthesis methods, the community must re-evaluate how experimental methods are written and communicated, to facilitate reproducibility, clarity, and text-mining accessibility.

But what is the current status quo for the writing of synthesis methods? Has the canonical writing style changed over time? Do researchers in different fields write with different styles? To shed light on these questions, we use previously-developed methods for text mining the literature [1] to measure the lexical complexity (i.e., normalized unique vocabulary size) of synthesis recipes with respect to material categories, the year of publication, and the number of times an article has been cited. We find that the lexical complexity of recipes is essentially invariant with respect to all these factors, as shown in Figure 1. Although new materials, synthesis methods, and lab equipment have proliferated over the decades, the fundamental vocabulary used for describing scientific experiments has remained static. This agrees with our anecdotal findings from annotating thousands of materials synthesis methods across hundreds of journals. It would appear that, in describing laboratory materials syntheses, implicit norms have resulted in a homogenized or prototypical writing style: "The precursors were purchased, the materials were mixed and heated, and finally the product was obtained."

Yet, few experimental methods describe materials synthesis in a clear, literal, and linear manner: comments on optimal equipment settings, observations about passive or observed events (such as color changes), and remarks on intermediate results (product yields, morphologies, chemistries, etc.) are ubiquitous and interwoven with actions taken by the researchers, such as heating or mixing. Experimental methods commonly introduce abridged, summarized, or non-linear procedures as well: "the above steps were repeated except the heating was performed at 500 °C." A canonical style for synthesis methods clearly exists, but it must often be distilled into a precise ontology by the reader using a surprisingly high level of inference.

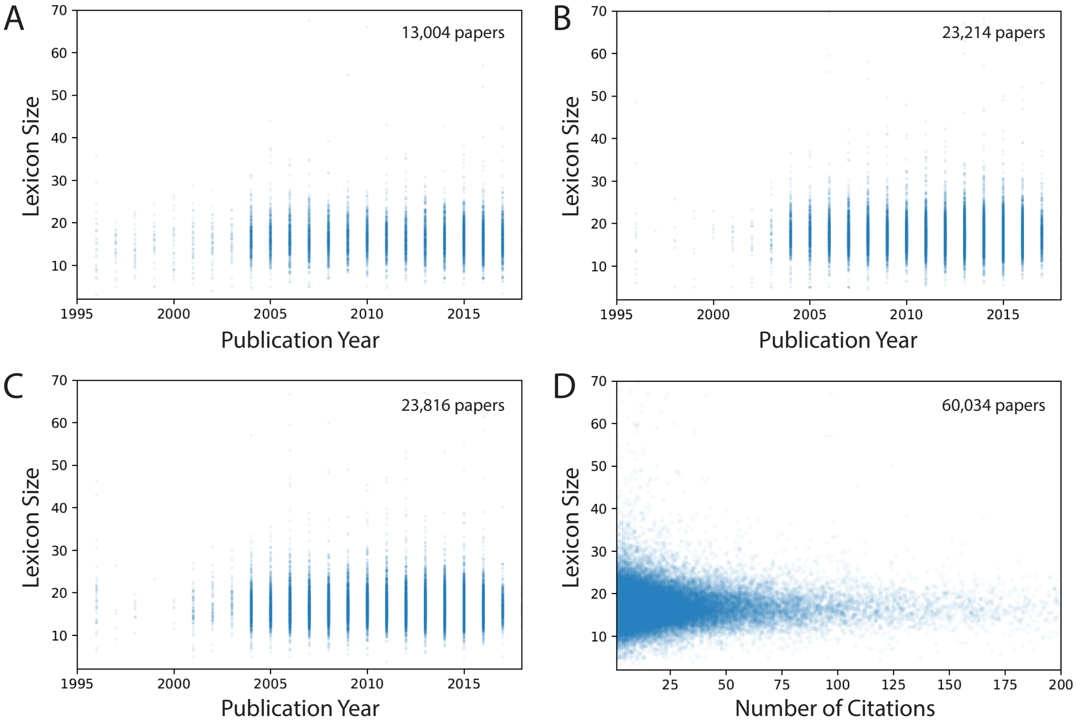


Figure 1. Lexical complexity (average unique words per paragraph, normalized by number of sentences) of various inorganic syntheses as reported in the literature. (**A**) Normalized lexicon sizes for recipes from journal articles with "perovskite" in the title. (**B**) Normalized lexicon sizes for recipes from journal articles with "nano" in the title. (**C**) Normalized lexicon sizes for recipes from journal articles with "cathode" in the title. (**D**) Normalized lexicon sizes for recipes from all aforementioned journal articles, plotted against the number of citations each article has received.

To further illustrate this point, we consider a hypothetical example motivated by materials science literature. Suppose the following two sentences are encountered in the experimental methods section of a journal article:

> The sample was prepared by a hydrothermal route using the precursors as received. The precursors were first dissolved in deionized water and then placed in a sealed autoclave at 200 °C for 10-12 h.

While some key details of the synthesis are clear, such as the reaction temperature and the type of synthesis being performed, resolving the finer details of such a synthesis requires high-level inference which may be onerous for machine-driven approaches. Clarifying the nature of the ’the sample’ may simply require scanning the article for a definition – and while this is easy for humans, this sort of distant co- or cross-referencing problem is comparatively challenging for machines, as the chance of finding false positives is high. This is particularly acute when trying to resolve the material being synthesized, as the full chemical specification of the target material is often not included in the methods section itself, having already been named in the abstract or introduction.Likewise, understanding the nature of ambiguous reaction conditions, such as “10-12 h,” often requires either domain knowledge or additional context from the article. Depending on the synthesis method, it may be that precise dwell times are unimportant to the result, or that various dwell times were reported (e.g., to make multiple samples). Finally, resolving the intent of lab actions often requires an internal model of the physical world. This is trivial for human experts – provided that they have the requisite domain knowledge – but it is formidable task for a machine for infer that when a material is “placed in a sealed autoclave,” it is being heated under autogenous pressure.

Given the level of prior knowledge needed to understand a synthesis route, is this style of scientific writing still deserving of the status quo in an age where scientific literature has an ever-expanding audience? Many scientists read and publish literature in English, regardless of whether or not it is their first language. Moreover, text-mining algorithms are increasingly being used to aggregate and understand experimental data at large scales [1,3]. If the ultimate goal of reporting experimental methods is to enhance transparency and reproducibility, then the community ought to strive for a writing style that maximizes comprehension for all readers, human and machine alike.

A key step in improving the understanding, transferability, and communication of experimental methods is to impose a canonical *ontology* for materials synthesis. Following in the footsteps of the well-established Gene Ontology [5], a materials synthesis ontology should consist of a controlled vocabulary with restricted relations between concepts.

In our proposed ontology, we consider a controlled vocabulary consisting of named entities: materials, operations (i.e., actions performed by experimenters), numbers, units, unit types (e.g., temperature), apparatuses, descriptive words (e.g., powder), and reaction conditions. Some of these entity types may be linked to one another in a specific fashion: for example, amounts may only be linked to materials, and reaction conditions may only be linked to operations. The key detail in our ontology is that the “backbone” of a synthesis is a linked chain of in-lab operations. For example, a typical solid state synthesis route may contain the operations, “mix, grind, sinter, cool.” The structure of such an ontology implies that, at the highest level of abstraction, the critical information to communicate is the precise sequence of actions that an experimenter performed on the materials.

Given the backbone of operations linked in a sequence, the rest of the synthesis route is hierarchically associated with the operations. Materials are linked to operations to denote which materials were acted upon in each experimental step. Detailed attributes of materials (e.g., amounts) and reaction conditions (e.g., temperatures) are linked to materials and operations, respectively, by connecting the appropriate numbers and units. Reaction conditions and apparatuses are linked directly to their relevant operations. An example of this ontology applied to a literature-excerpted synthesis route is shown in Figure 2. Additionally, we provide the full details for this ontology schema at www.synthesisproject.org.

The application of a synthesis ontology across numerous journal articles aids in understanding patterns in the literature. The text-mined synthesis data used to produce Figure 1 also contains underlying operation sequences, where in-lab actions are sorted by the order of text appearance. We sort these operations into their most common subsequences by brute force calculation. Both “cathode” and “perovskite” syntheses contain ubiquitous operation subsequences such as “filter, wash, dry”. However, the “cathode” syntheses frequently contain the subsequence, “mix, coat, dry,” while the “perovskite” syntheses frequently contain “calcine, press, sinter.” These operation subsequences are characteristic of common synthesis methods relevant for each category of materials.

While this ontology captures many details of written experimental methods, some types of information cannot be placed easily and unambiguously into this framework. Operations written in non-chronological order and cross-references to experimental conditions or materials across different paragraphs are among the most challenging pieces of data to canonicalize. Even with the assertion of a precise synthesis ontology, we are still left with significant challenges in annotating existing and future written experimental methods.

What, then, is required of an ideal written experimental procedure? We have proposed that ideal synthesis methods should be written in a way to facilitate the rapid and unambiguous inference of well-defined *synthesis ontologies*. In other words, we argue that a written synthesis method should allow a reader (human or machine) to easily infer from a synthesis method details such as "heat, TiO_2, 800 K, 2 h." This suggests immediate changes to the status quo for writing synthesis methods.

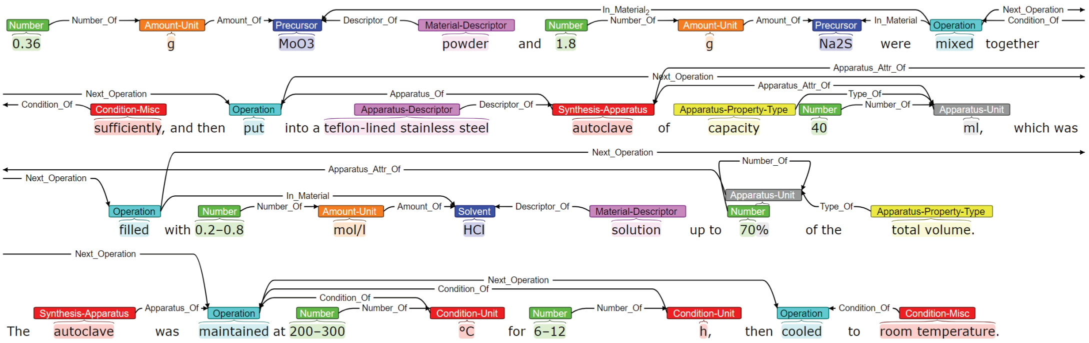


Figure 2. A prototypical annotated synthesis excerpt [6], demonstrating the application of our synthesis ontology to a written experimental method. Colored blocks of text represent named entity labels, and arrows denote relations between entities. Annotation of this text was performed using the BRAT annotation software (brat.nlplab.org).

First, synthesis methods must avoid ambiguous entities, including materials and quantities and second, the canonical writing style for synthesis methods must be significantly restructured to better delineate individual events. Further, there must be a clear separation between the description of the methods and experimental observations, as the latter can be described within the results and discussion sections of an article.

We first discuss the issue of ambiguous quantities. While all measurements have inherent uncertainty, the current body of literature is rife with unspecified synthesis conditions. For example, writing “8 hours” instead of “overnight” lessens the burden of readers’ inference. Critically, even if “8 hours” is an estimated value, writing the estimated value in the original synthesis method reduces inconsistent estimations on the parts of the readers (whereas “overnight” may be interpreted differently across readers).

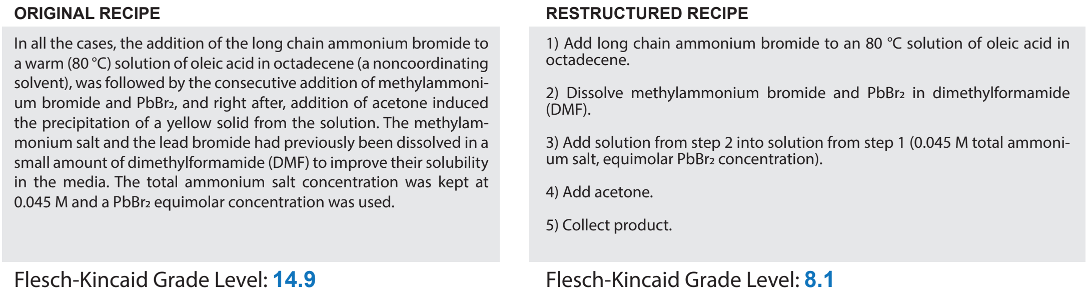
While this insistence on precise communication may seem cumbersome at first, it is critical for both improving reproducibility at scale, and also for accelerating the development of automated data-driven synthesis techniques. For example, robot-driven syntheses [7] require precise instruction sets with well-defined quantities. Additionally, aggregate data mining of reported synthesis parameters is only tractable for numerically-reported quantities - otherwise, data imputation must be performed by imposing assumptions across experimental methods.

Similarly, all materials which are known to the researcher (i.e., their full chemical specifications are known) should be explicitly written out as such. For example, rather than referring to precursors or target materials with abbreviations, by sample names, or by generic material classes, full, standardized chemical names or formulas should be used for each material when available. Moreover, when multiple, chemically similar target materials are synthesized, the chemical specification for each should be written out, rather than abbreviated. For instance, if multiple metal diselenides were synthesized, a list of fully specified materials ($TiSe_2$, $MoSe_2$, and WSe_2) should be favored over abbreviated forms of reporting (MSe_2, M = Ti, Mo, W), despite the added length. Doing so relieves the reader, particularly machines, from having to accurately cross reference these mentions with their proper antecedents elsewhere in the article text. By mitigating this ambiguity, the likelihood of text extraction or experimental laboratory error may be commensurately reduced.

Even if given an ideal synthesis method, with all numerical quantities clearly stated and no ambiguous phrases used, we would yet advocate for a significant restructuring of the writing style. The prototypical experimental methods section is written in past-tense, using passive voice, with an impartial tone to describe the steps carried out during an experiment. The synthesis section of a materials science article would never dictate, "Heat TiO_2 at 800 K for 2 h," as if excerpted from a cookbook. However, based on our experience from text-mining millions of journal articles and manually annotating thousands of materials synthesis methods, this “cookbook”-style language is exactly what is needed for machine-readability. Each “instruction” is presented as an imperative, present-tense sentence which states only the relevant details for the current synthesis action. Surprisingly, such a structure not only preserves the human-readability of a synthesis route, but improves it.

Figure 3 shows an example synthesis method along with a restructured version in this new format. The Flesch-Kincaid grade level score [8], corresponding roughly to the US grade level education required to understand the text, is lowered substantially in the restructured version. By omitting complicated verb tenses, authors can avoid unnecessarily confusing readers. Moreover, the line-by-line segmentation of each synthesis action vastly reduces the complexity of text-mining, as each synthesis action is trivially found at the beginning of each new line, and any materials associated with an action appear within the same line.

Beyond these simplifications for the reader, this imperative step-by-step style enforces a strict temporal order on the steps used during the synthesis of a material. The additional structure provided by this writing style may ultimately improve reproducibility for the community at large, as the experimenter is forced to write the explicit narrative of all actions that were done in the lab, similar to the recordings in lab notebooks. Moreover, a line-by-line structure for reporting experimental methods will vastly simplify the process of comparing across published synthesis routes. As an example, some inorganic compounds have seldom-synthesized metastable phases (e.g., brookite titania) and the synthetic conditions that select for these phases are unclear. Canonicalized reporting of the experimental methods would vastly simplify the process of detecting outliers in future syntheses.



ORIGINAL RECIPE

In all the cases, the addition of the long chain ammonium bromide to a warm (80 °C) solution of oleic acid in octadecene (a noncoordinating solvent), was followed by the consecutive addition of methylammonium bromide and $PbBr_2$, and right after, addition of acetone induced the precipitation of a yellow solid from the solution. The methylammonium salt and the lead bromide had previously been dissolved in a small amount of dimethylformamide (DMF) to improve their solubility in the media. The total ammonium salt concentration was kept at 0.045 M and a $PbBr_2$ equimolar concentration was used.

Flesch-Kincaid Grade Level: **14.9**

RESTRUCTURED RECIPE

1) Add long chain ammonium bromide to an 80 °C solution of oleic acid in octadecene.

2) Dissolve methylammonium bromide and $PbBr_2$ in dimethylformamide (DMF).

3) Add solution from step 2 into solution from step 1 (0.045 M total ammonium salt, equimolar $PbBr_2$ concentration).

4) Add acetone.

5) Collect product.

Flesch-Kincaid Grade Level: **8.1**

Figure 3. An example plain-text synthesis recipe excerpt [9] and a restructured synthesis recipe written in imperative present tense, along with each of their respective Flesch-Kincaid grade-level scores [8].

Nevertheless, a stark shift in the way that experimental methods are written is bound to be met with resistance. But the status quo has failed us: there is undoubtedly a reproducibility crisis across nearly all fields of scientific research [10]. Asserting a synthesis ontology and actively changing writing styles are necessary short-term efforts in order to improve the cohesiveness of these fields.

It is a tall order to suggest a departure from a canonical writing style that has persisted for decades. Nonetheless, we believe that the time for improving the communication of materials synthesis methods is now. Major journals have already begun to prioritize a focus on methods writing [11], and machine-guided synthesis has recently become a physical reality [7]. Closing the gap between human-readable and machine-readable methods will extend the impact of the insights contained in each published synthesis method and contribute towards a global body of unified materials synthesis knowledge.

Acknowledgments: We would like to acknowledge funding from the National Science Foundation Award #1534340, DMREF that provided support to make this work possible, support from the Office of Naval Research (ONR) under Contract No. N00014-16-1-2432, and the MIT Energy Initiative. Early work was collaborative under the Department of Energy’s Basic Energy Science Program through the Materials Project under Grant

No. EDCBEE.The authors would also like to thank Dr. Lee Cronin for providing the inspiration to compile our thoughts on the topic.

Conflicts of Interest: The authors declare no conflicts of interest.

Bibliography

1. Kim, E.; Huang, K.; Saunders, A.; McCallum, A.; Ceder, G.; Olivetti, E. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chemistry of Materials* **2017**, *29*, 9436–9444.
2. Segler, M.H.; Preuss, M.; Waller, M.P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604.
3. Swain, M.C.; Cole, J.M. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *Journal of Chemical Information and Modeling* **2016**.
4. Raccuglia, P.; Elbert, K.C.; Adler, P.D.; Falk, C.; Wenny, M.B.; Mollo, A.; Zeller, M.; Friedler, S.A.; Schrier, J.; Norquist, A.J. Machine-learning-assisted materials discovery using failed experiments. *Nature* **2016**, *533*, 73.
5. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; others. Gene Ontology: tool for the unification of biology. *Nature genetics* **2000**, *25*, 25.
6. Li, W.J.; Shi, E.W.; Ko, J.M.; Chen, Z.z.; Ogino, H.; Fukuda, T. Hydrothermal synthesis of MoS2 nanowires. *Journal of Crystal Growth* **2003**, *250*, 418–422.
7. Granda, J.M.; Donina, L.; Dragone, V.; Long, D.L.; Cronin, L. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* **2018**, *559*, 377.
8. Kincaid, J.P.; Fishburne Jr, R.P.; Rogers, R.L.; Chissom, B.S. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel **1975**.
9. Schmidt, L.C.; Pertegtás, A.; González-Carrero, S.; Malinkiewicz, O.; Agouram, S.; Mínguez Espallargas, G.; Bolink, H.J.; Galian, R.E.; Pérez-Prieto, J. Nontemplate synthesis of $CH_3NH_3PbBr_3$ perovskite nanoparticles. *Journal of the American Chemical Society* **2014**, *136*, 850–853.
10. Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature News* **2016**, *533*, 452.
11. Marcus, E. A STAR Is Born. *Cell* **2016**, *166*, 1059–1060.