Advances in Cognitive Systems X (20XX) 1–6 Submitted X/20XX; published X/20XX

Cognitive Modeling of Paraphrase for Authorship Anonymization

Marjorie McShane MARGEMC34@GMAIL.COM
Sergei Nirenburg ZAVEDOMO@GMAIL.COM
Christian Arndt CHRIS.ARNDT.18@GMAIL.COM
Sanjay Oruganti SANJAYOVS.RPI@OUTLOOK.COM
Jesse English DRJESSEENGLISH@GMAIL.COM
Rensselaer Polytechnic Institute, Troy, NY, 12180, USA

Abstract

We report an explanatory cognitive model of paraphrase and its implementation in a system for automatic authorship anonymization. The model covers what it means to faithfully retain *meaning and discourse coherence* in a paraphrase, how do deal with *polysemy* given that full semantic analysis of open text is beyond the state of the art, and how to define and characterize an author's *style*. We also discuss how this knowledge-based approach can be integrated with large language models in neurosymbolic systems that that combine the coverage of LLMs with the reliability and explanatory power of symbolic processing.

1. Introduction

Authorship anonymization involves automatically paraphrasing texts to retain their meaning while making it impossible for stylometry systems to identify the author or salient characteristics of the author.[1] Other names for it are author obfuscation, adversarial stylometry, and privacy protection. Authorship anonymization can have prosocial applications, such as protecting the identity of whistleblowers, authors writing under a pseudonym, and reviewers. It can also have antisocial applications, such as hiding scammers, people spreading disinformation, and writers of fake reviews.

When people paraphrase, they orient around the meaning they want to express. Machines can't take this approach because full semantic analysis of open text is beyond the state of the art. This leaves the following three options:

1. Use machine learning (ML): One can sidestep the need to compute meaning by using ML to paraphrase (e.g., McDonald et al., 2012; Bevendorff et al., 2019). Most recently, ML-based paraphrasing is being carried out by large language models (LLMs), which can be configured to paraphrase individual sentences or multi-sentence chunks of text.[2] Paraphrasing larger

[1] For background on stylometry, see Abbasi and Chen (2008). For a nice graphic (their Fig. 1) showing how text analysis and synthesis overlap with paraphrasing, see Burrows et al. (2012), who report work on using crowdsourcing and machine learning to compile a corpus of paraphrases.

[2] This is being pursued by collaborators on the project funding this research; see the Acknowledgments.

chunks of text can generate texts that are quite different from the original, thus fulfilling the goal of anonymization. However, paraphrasing by LLMs – like all work by LLMs – is unreliable. For example, for the input "Because he was impatient, his subordinates hated having to work with him," one LLM we experimented with offered the paraphrase "He was disliked by his subordinates because he was too hasty". This is not a felicitous paraphrase one two counts: *hasty* is a rare word in modern English, which turns a stylistically neutral sentence into one that sounds unnatural; and *hasty* is semantically quite different from *impatient,* with *impatient* more clearly implying that his behavior directly affected his subordinates.[3]

2. Use knowledge-based modeling: One can compile an inventory of paraphrases – strings, open patterns, and syntactic transformations – that can reliably replace each other in any context, without the need for full semantic analysis, and then implement a system to carry out those replacements. This approach is the one detailed in this paper. It is grounded in a cognitive model that must account for what it means to faithfully retain *meaning and discourse coherence* in a paraphrase (section 1.1), how to deal with lexical *polysemy* given that the system cannot rely on a full semantic analysis of the text (section 1.2), and how to define and characterize an author's *style*, which should be markedly different in the source and paraphrased versions of the text (section 1.3). The downside of this approach is that the size of the paraphrase repository determines the extent to which the text will be anonymized, and building that repository requires resources, which are always in short supply.
3. Use a neurosymbolic approach: One can combine knowledge-based and LLM-based capabilities into a neurosymbolic system that aims to optimize what each can deliver. We have implemented certain aspects of such hybridization in the system we report, and we sketch work in progress that involves additional LLM-based contributions.

Our approach to anonymization involves the following steps:

1. Paraphrase the text using knowledge-based methods, which support strict paraphrasing, explainability, and relatively high reliability, albeit currently limited coverage.
2. Vet all paraphrased sentences using an LLM to weed out unusual formulations. This is exactly the kind of thing that LLMs are best suited for since it involves the statistical likelihood of sequences of strings (Mahowald et al., 2023; Bubeck et al., 2023). For example, our paraphrase inventory erroneously included the bidirectional equivalents *very* and *really*, which are not

[3] A loose definition of paraphrase is also used in the Microsoft Research Paraphrase Corpus, which contains 5801 sentence pairs that were hand-labeled to indicate whether or not the pair constituted a paraphrase. But, as Dolan and Brockett (2005) write, the paraphrases in that corpus actually reflect a "relatively loose definition of semantic equivalence." For example, they say that "any 2 of the following sentences would have qualified as 'paraphrases', despite obvious differences in information content:

The genome of the fungal pathogen that causes Sudden Oak Death has been sequenced by US scientists
Researchers announced Thursday they've completed the genetic blueprint of the blight-causing culprit responsible for sudden oak death
Scientists have figured out the complete genetic code of a virulent pathogen that has killed tens of thousands of California native oaks
The East Bay-based Joint Genome Institute said Thursday it has unraveled the genetic blueprint for the diseases that cause the sudden death of oak trees".

interchangeable in all contexts and, therefore, should not have been included without further contextual constraints: e.g., *Do you really believe… / *Do you very believe…* An LLM can readily detect such problems, allowing the anonymizer to reject such paraphrases.

3. In all cases where the knowledge-based system can offer multiple paraphrases, use an LLM to select among them.[4] For example, our system paraphrased the following sentence in three ways, and several LLMs we tested selected the second as the best.[4]

   Drivers are not required to record traffic hazards, though, and apparently **seldom** do so.

   1. **Drivers are not bound to** record traffic hazards, though, and apparently **hardly ever** do so.
   2. **Drivers do not need to** record traffic hazards, though, and apparently **hardly ever** do so. ✓
   3. **Drivers have no need to** record traffic hazards, though, and apparently **hardly ever** do so.

   However, the LLM's preference should not necessarily always be selected since it might be the one that least modifies the original text. So, a post-LLM selection process should introduce some variability into the ultimate selection.
4. Using an authorship attribution system, determine if the degree of obfuscation achieved by the knowledge-based system is sufficient to anonymize authorship. During system development, this can be estimated, but results might differ across authors, topics, genres, and actual texts.
5. If the knowledge-based system does not adequately anonymize texts, then have an LLM-based anonymization system treat some percentage of the as-yet untouched sentences. That percentage should be as low as possible while ensuring anonymization since the LLM is prone to change the meaning of texts and its actions cannot be explained. As with stage 3, this needs to be optimized through testing.

Two points of clarification are in order at this point. First, readers acquainted with our research lab's main line of R&D – configuring Language-Endowed Intelligent Agents (LEIAs) within the OntoAgent content-centric cognitive architecture (Nirenburg, McShane, & English, 2020; McShane, Nirenburg, & English, *forthcoming*) – should note that the work reported here does not leverage the full suite of capabilities of LEIAs. This is because authorship anonymization requires coverage of open text, and deep semantic analysis of open text is beyond the state of the art. However, the work does leverage resources and capabilities developed for LEIAs, including an ontology, a computational lexicon, syntactic analysis, morphological analysis, various language processing submodules, and decades' worth of experience in computational semantics. Repurposing these resources for a practical, quick ramp-up application is an example of how results of a long-term program of R&D can be useful in the near term, which is a practical necessity. It is also noteworthy that this work has deepened our understanding of how polysemy and paraphrase can be dealt with computationally, which feeds back into LEIA development.

[4] We use an LLM in a similar manner for text generation by our cognitive agent system (McShane, Nirenburg, & English, forthcoming).

[4] The LLMs used were Claude 3 Opus, Claude 3 Sonnet, Claude 3 Haiku, GPT-4, GPT-3.5, and Gemini Advanced. The prompt was: "From each group below, select the sentence that demonstrates correct English grammar, syntax, and structure, without providing explanations for the choices. Give me the option number for each set of choices in the form of a table."

The second point of clarification is that this work had to conform to externally decreed rules that did not permit us to implement all of the above workflow —which begs a separate discussion of the effects of external pressures on scientific inquiry that we cannot open up here.[5]

To date, we have implemented and informally evaluated the above algorithm through step 3. We are currently working on step 5. We are investigating opportunities to get access to a system for step 4 to enable overall system optimization.

1.1 Faithful Retention of Meaning and Discourse Coherence

We understand *faithful retention of meaning and discourse coherence* to mean that only the surface form of the text can change: no information can be added, removed, or modified, and the sentence must continue to sound like normal English in its larger context. (1) – (4) are examples of true paraphrases under this definition.

(1) a. **To** do well, you have to study hard. b. **In order to** do well, you have to study hard.
(2) a. **Apart from** him, nobody else came. b. Nobody else came, **except for** him,
(3) a. This **led to** a big debate. b. This **resulted in** a big debate.
(4) a. **It goes without saying that** this was the **right** decision. b. **Clearly,** this was the **correct** decision.

Although a passing acquaintance with thesauri and wordnets might give one the impression that language is bursting with synonyms, most of the entities clustered in such resources are not synonyms in the strict sense – they are at best plesionyms, words that are semantically related in a large variety of ways. The function of such resources is to jog writers' memories when they are trying to recall the precise word that is needed for a particular context. This means that one cannot just replace one word with something listed as a synonym in online lexical resources and expect to retain the meaning and/or fluency of the text. For example, thesaurus.com lists the following as the closest synonyms of *student*: *graduate, undergraduate, junior, pupil, scholar*. Replacing in either direction leads to errors. For example, one cannot replace *student* with *undergraduate* because not every student is an undergraduate; and if the text contains *undergraduate student* then replacing *undergraduate* with *student* would yield *student student*.[5]

The need to retain discourse coherence means that syntactic transformations cannot be randomly applied.[6] For example, *Charlotte fixed the fence* cannot be subject to the following transformations unless it is warranted by the discourse context: [subject dislocation] *Charlotte, she fixed the fence*;

[5] According to the requirements of the research program within which this research was conducted, accessing an LLM through the Cloud for steps 2 and 3 is out of bounds. However, we do not have the resources to implement a local LLM. So, we are using the Cloud for the system we report here but cannot use this system as a deliverable for the project. We also do not have access to an authorship attribution system, so we cannot experiment with step 4.

[5] For further discussion of the use of thesauri and other human-oriented resources for developing computational-linguistic systems, see McShane, Nirenburg, and English (2024, forthcoming).

[6] Some types of paraphrase that have been identified in the linguistic literature (e.g., Bhagat & Hovy, 2013) have not yet been included in the system but are on agenda: e.g., the expression of social roles (*Fred is a first-grade teacher ↔ Fred teaches first grade*) and the expression reported speech (*John said, "I think I'll attend" → John said he thought he would attend*).

[object dislocation] *Charlotte fixed it, the fence*; [it-was topicalization] *It was Charlotte who fixed the fence*; [as-for topicalization] *As for Charlotte, she fixed the fence*.

Although these variants retain the basic meaning of the original sentence, using them to replace the active form in a particular context is likely to result in either a disruption to the discourse structure or the addition of a new meaning. For example, passivizing sentences allows the theme (topic) to occupy the subject position, thus linking the new sentence to the preceding context. So, one cannot randomly passivize and unpassivize sentences and expect them to retain discourse coherence. Similarly, the dislocation and topicalization structures above draw special attention to particular arguments in a way that would disrupt the flow of the text if such emphasis was not warranted.

Although it is important to not alter the meaning of the original text, we might want to permit certain kinds of stylistic infelicity in service of obfuscation. For example, our informal experiments included the following paraphrases which, depending on one's evaluation criteria, might be considered acceptable or not acceptable.

(5) I knew that there **is** a possibility that that might make her more likely to speak, and I still did it because I had to limit the contact. [is should be was]
(6) Encouraged, I told the nurses to leave her off the machine indefinitely, with the idea that there **is** a chance that she might go the whole night unassisted. [is should be was]
(7) Therefore, negative results give an **untrue** sense of security if they are interpreted as meaning that the product is free of the microorganism sought. [the original false sense of security is idiomatic]
(8) The community would take those kids away and do the job for them if families were so irresponsible as to fail to educate their children! [when the clause order what switched, the sequence of coreferential expressions became not ideal]
(9) Her health … has kept her at home, where Harry could **most of the time** find her. [*most of the time* replaced *usually;* ideally, it would either have commas around it or would be at the end of the sentence]

Comparing paraphrases to original texts is similar to reading texts that you know are a translation from another language: it is natural to be hyperaware of, and even question, stylistic choices. But if you find those same choices in a native-language text, you don't think twice.

The paraphrases above would make the author's style less academic, which might be a valuable obfuscation strategy. In terms of operationalizing it, one could introduce rules, for example, to disrupt canonical sequences of tenses, which some highly accomplished non-native speakers – and even some native speakers – do not consistently use according to prescriptive norms.

Evaluations of automatic paraphrasing should avoid inadvertently reflecting the evaluators' idiolects, stylistic preferences, or notions about prescriptive grammar.

1.2 Lexical Polysemy

Most words and many multiword expressions in any language are polysemous. Any given *sense* of a word or expression might have a close synonym that could result in a strict paraphrase, but identifying that sense requires semantic analysis. For example, *country* can be paraphrased by *nation* in some contexts but not in the sentence *He lives in the country, far away from the city*.

An inroad to dealing with lexical polysemy is to focus on the construction-based nature of languages. That is, language is constructed not of wholly compositional words but, instead, of constructions made up of combinations of words, punctuation marks, and/or variable slots.

Paraphrasing requires understanding which components of sentences are acting as units and then determining whether that unit can be paraphrased. Identifying which multicomponent strings are linguistically useful targets for paraphrasing cannot be done automatically.[7]

An important finding from our work is that, for purposes of computational cognitive modeling, a broadly inclusive definition of *construction* is most useful (McShane & Nirenburg, 2021; McShane, Nirenburg, & English, 2024). Clearly, constructions cover the traditionally acknowledged inventory: idiomatic expressions (*take a load off*), non-idiomatic fixed expressions (*Have a nice day*), phrasal verbs (*buck* up), syntactic transformations (passivization, object fronting), and the like. However, for purposes of automatic paraphrasing—as well as for configuring the language understanding and generation components of agent systems—**constructions should include other multicomponent entities whose combination allows for disambiguation of the individual components**. For example:

- A word or expression can be reliably paraphrasable in a particular text position. For example, when the word *additionally* is used sentence-initially and is followed by a comma, it is a discourse connector that links the given sentence to the previous one and carries the meaning of elaboration. It can be paraphrased by several other expressions that also must be sentence-initial and followed by a comma, such as *in addition* and *moreover.* So, although *additionally* is a single word, its construction comprises three elements: sentence-initial position, the word itself, and the comma that follows.[8]
- A frequently-encountered sequence of words can be paraphrasable by a different sequence even though the individual words in isolation are not reliable paraphrases for each other. For example, *a couple of minutes* ↔ *a few minutes* and *It's not what it looks like* ↔ *It's not what it appears to be* are reliable paraphrases even though the words *couple/few* and *look like/appear to be* are not interchangeable in all contexts.
- A word or expression can be replaceable by another one as long as it is preceded or followed by something specific – be it a word from a list, a syntactic constituent headed by a particular word, a word in a particular part of speech, or a particular kind of syntactic constituent. For example:
  - *concerned with* and *concerning* are paraphrases as long as they are preceded by a noun phrase headed by the word *issues, studies, questions, theory,* or *approach*.
  - *It's a further* can be paraphrased by *It's another* as long as they are followed by a noun phrase headed by the words *reason, example, sign, thing, opportunity, attempt, problem, piece, study, reminder, step, indication, question, complication, increase, decrease,*

[7] For example, to compile their list of 505 useful phrases for language pedagogy, Martinez and Schmitt (2012) had to manually prune an automatically generated list of n-grams. Their report includes a nice overview of the literature about automatically creating multiword expression lists.

[8] For clarity of presentation, we are not including all variations of constructions presented as examples. For the case above, *additionally/in addition/moreover* can also follow certain other punctuation marks, such as a semi-colon, and they might not be followed by a comma. However, as one moves away from the most canonical situation, the reliability of pattern identification and substitution tends to drop. For example, if *additionally* is preceded only by a comma and not followed by any punctuation, then it might not represent the paraphrase set we're talking about, as in the COCA corpus example "GWAS, additionally known as whole genome association studies, is a genome-wide approach…"

development, dimension, challenge, limitation, blow, distinction, refinement, argument, consequence, or *delay*.
- *Look*$_V$ *to Pronoun for* can be paraphrased by *consult Pronoun for* as long as they are followed by a noun phrase headed by *help, leadership, guidance, support, answers, assistance, encouragement,* or *inspiration*.
- *Bring up* can be paraphrased by *raise* as long as their direct object is headed by *topic, issue, subject, fact, question, idea, point, matter,* or *possibility.*

In reading these examples, you are likely to have noticed several things:

- Currently, the word lists associated with constructions are incomplete, having been compiled through introspection and online search of the COCA corpus. They can be expanded given more time. But such word lists are essential, since *not* constraining the constructions to include the words in the lists would lead to incorrect paraphrases.
- The elements of some lists fall into semantic classes. This observation is useful for acquisition of an ontological-semantic lexicon of the type we are developing for agent systems. However, for the anonymization project, listing – albeit incomplete – is the only approach we have time for.
- The examples might look like the beginning of a potentially endless list of frequent expressions in English. This is not far from the truth, but it does not invalidate the approach. For purposes of anonymization, paraphrasing need not address every single text component, only enough to mask the author,[9] which can only be assessed by testing the output of author anonymization systems using author identification systems.
- The choice of what to consider a variable versus a constant can be tricky but, for the current purposes, it is based on human judgment. For example, *It's a further / It's another* + [*reason, example,* etc.] treats *It's* as a constant. There is a separate construction for *That's a further / That's another* [*reason, example*, etc.].

Moving from practice to theory, we think that it is psychologically plausible that people store these kinds of constructions in their mental lexicons. This is why native speakers of English are likely to come up with very similar sets of paraphrases for given sentences (something that could be tested in psycholinguistic experimentation[10]). For example, given the input *Not too long ago I changed jobs,* the chunk *not too long ago* can be paraphrased by *not long ago, just recently, recently, a short time ago* and *a short while ago*; the chunk *I changed jobs* can be paraphrased by *I switched jobs*, *I got a new job,* and *I left my old job for a new one*; and a comma after the sentence-initial adverbial is optional. All of these paraphrase opportunities create a large set of strict paraphrases from which the obfuscation system can select.

Importantly, paraphrases can be reliable in one direction but not the other. This can occur for various reasons. For example, it can be fine to paraphrase using a slightly more generic term – *policewoman → police officer* – but not the other way around, since not every police officer is a

[9] Similarly, when building cognitive systems, the acquisition of expressions can be guided by the domain covered by a particular application.

[10] A relevant direction of research involves how construction frequency interacts with memory (e.g., Bybee, 2013: p. 49).

policewoman. Similarly, an unambiguous word or multiword expression can be paraphrased by an ambiguous one but not the other way around: *waitress → server* but not *server → waitress* (the server in the context might be a male person or a computer device). The judgments about "slightly more generic" and directionality of confident paraphrases must be made by people.

Finally, there are standard ways of saying things, and switching out components of a canonical expression can result in unnatural formulations. For example, replacing *I would appreciate it if you would…* with *I would value it if you would…* sounds unnatural, even though it is grammatical and understandable. Similarly, although changing the ordering of adjectives would lead to meaning-preserving modifications of texts, adjective order is not random. It must follow the so-called royal order of adjectives which dictates, for example, that *tall and handsome* is correct whereas *handsome and tall* is not. In some cases, multiple adjectives within a given category have a preferred ordering, whereas in others, different orderings are acceptable. To generalize, languages consist of *normal ways of saying things* that native speakers memorize. When non-native speakers—or computer programs manipulating texts—get their point across with sentences that sound unnatural, they are straying from the norm in ways that would be easily detectable by any native speaker.

One challenge in manually acquiring reliable paraphrase alternations is that untrained people have a hard time recognizing polysemy. In a resource-limited development effort, acquirers have to make split-second decisions about whether the source variant has any meanings or uses that would not be correctly paraphrased in all contexts by the target variant. An informal experiment with undergraduate students suggested that they struggled to detect ambiguity, so only about a third of their suggested paraphrases proved useful. The lion's share of the paraphrase database was compiled by the first author without only minimal consultation of text corpus evidence. Relying more heavily on a corpus would have been the most reliable, but prohibitively expensive, way to carry out the work. We use an LLM to vet the felicity of all paraphrases generated by the anonymization system on the basis of the paraphrase database.

1.3 Defining and Characterizing an Author's Style

Our approach to changing the style of a text in order to anonymize it does not involve a literary scholar's notion of style or the transformation of a plain description of a sports match into the metaphor-infused language of sports commentators.[11] Instead, we define *stylistic features* as semantic and pragmatic features for which unambiguous paraphrases can serve as values. Each time an author uses one of the paraphrases in our database (e.g., *quickly* versus *rapidly*), this reflects a stylistic choice about how to convey that meaning. The sum of an author's choices between available paraphrases is the author's style; it is a list, not a descriptor.

To paraphrase the above: (1) The paraphrase correspondences in our database reflect **meanings** because they are unambiguous: no matter the context, they have a predictable meaning. By contrast, most words and many multiword expressions are not unambiguous outside of context so they cannot be included in the database. (2) For each expression in the database, there is at least one paraphrase. So, **speakers and writers of English have a choice when expressing this meaning**. (3) The speaker's/writer's **preference** for how to express this meaning **is a stylistic**

[11] Bevendorff et al.'s (2019) claim that "stylometry [is not] understood well enough to compile rule sets that specifically target author style" (p. 1098) is unfounded, relying on an unnecessarily narrow definition of style.

feature. (4) Every time the writer uses one of the expressions in our database, that **choice** means that the writer is not choosing the alternative option. (5) The **inventory of choices** when expressing the meanings in the paraphrase database are the **author's profile** – or, more specifically, the aspect of the author's profile that we can capture using this method at this stage of developing the paraphrase database.

In order to make the results of anonymization explainable, we name the stylistic features in our database and append these names as metadata to the automatically generated paraphrases. In some cases, a feature name follows conventional terminology: for example, active/passive. In other cases, the feature uses the name of the ontological concept that grounds the meaning in the OntoSem ontology: e.g., EXPRESS-EMPHASIS. And in still other cases, a proxy label is used that will suffice until such time as we have time to expand the OntoSem lexicon and ontology to accommodate all meanings covered in the paraphrase inventory. For example, ProxyAdv:[inherently,intrinsically] states that there is some meaning, as yet to be recorded in the ontology, that is shared by the adverbs *inherently* and *intrinsically*. Table 1 shows examples of features showing all three feature-naming conventions.

Table 1. Examples of feature labels and values showing all three explanatory naming conventions. Note that paraphrase sets can be strings, variable-inclusive patterns, or transformations.

Feature Label	Value (paraphrase sets)
Conventional linguistic function	
active/passive	Subj V DirectObj ↔ $Subj_{UnderlyingDirectObj}$ *be* $V_{PastPart}$ by $NP_{UnderlyingDirectObj}$
subj ellipsis in clausal coordination	Subj1 CL1 and (Adv) Pro1 CL2 ↔ Subj1 CL1 and (Adv) ___ CL2
prep-part (particles that are homographous with prepositions) ordering	[for non-pronominal DirectObjs only] • carry around DirectObj ↔ carry DirectObj around • drag around DirectObj ↔ drag DirectObj around
compound/of-PP	• gas shortage ↔ shortage of gas • depression risk ↔ risk of depression
ordering of conjoined adjectives	• bright and lively ↔ lively and bright • calm and smooth ↔ smooth and calm
empty filler words	• essential ↔ absolutely essential • throughout ↔ all throughout • cameo appearance ↔ cameo
Ontologically grounded meanings	
REQUEST-ACTION (FORMALITY .5) (POLITENESS .5)	Would you VP? ↔ Could you VP? ↔ Can you VP?
REQUEST-ACTION (FORMALITY .5) (POLITENESS .7)	Would you please VP? ↔ Could you please VP? ↔ Would you kindly VP?
EXPRESS-EMPHASIS	To put a fine point on it, … ↔ To emphasize, … ↔ Importantly,
EXPRESS-AN-OPINION	I think (that) CL ↔ My feeling is (that) CL ↔ In my opinion, CL
Implicit Meanings: Proxy labels	
ProxyAdv	inherently ↔ intrinsically
ProxyManner-Fashion-Ly	in a coherent manner ↔ in a coherent fashion ↔ coherently
ProxyNoun	acquisition of ↔ acquiring of

ProxySubjV	this involves ↔ this entails
ProxyV	affects ↔ has an effect on
ProxyAdj	thorough ↔ extensive

This concludes the overview of the three theoretical components of the paraphrasing model: what it means to faithfully retain *meaning and discourse coherence* in a paraphrase, how to deal with *polysemy* given that full semantic analysis of open text is beyond the state of the art, and how to define and characterize an author's *style*. We now turn to the practical matter of compiling the database of paraphrase correspondences to seed the knowledge-based anonymization program.

2. The Paraphrase Database

We have semi-automatically created, and continue to expand, a large inventory of confidently replaceable words, phrases, open patterns, and syntactic transformations in what we call the paraphrase database. All paraphrase variants must be interchangeable in all contexts, irrespective of the surrounding text. The replacements can be specified as bidirectional (*maybe* ↔ *perhaps*) or unidirectional (*nation* → *country*). Knowledge acquisition strategies include the use of:

- linguistic knowledge bases – online thesauri, word lists, etc. – to jog the memory of acquirers
- the online search engine for the COCA corpus (Davies, 2008-), which is particularly useful for populating lists of words that constrain the variable slots in constructions
- GoogleTranslate for roundtrip machine translation (English → French → English) to suggest paraphrases that people may not think of including (listing is difficult for people whereas confirming the accuracy of something presented is not)[12]
- the online Diffchecker text comparison tool (https://www.diffchecker.com/text-compare/), which highlights the differences between the before and after variants of roundtrip translation and makes it much faster to manually compare them (this comparison cannot be done automatically)
- homegrown data analytics for exploring linguistic hypotheses about paraphrases in the COCA corpus
- results of our lab's past work on computational semantics, including our computational lexicon, ontology, an inventory of speech acts, constructions that convey them, and more.

Detecting context-independent paraphrase equivalents cannot be teased out of big data using machine learning, LLMs, or data analytics by themselves. Contrary to popular belief, manual knowledge acquisition has persisted in some guise even in the age of LLMs. Humans have been working on text annotation, cleaning the machine learning datasets, tagging the images that train vision recognition systems, and preparing application systems (like voice assistants) to respond in

[12] Roundtrip translation has been used as a method of evaluating machine translation systems.

specific ways to specific inputs. What differs between knowledge-based and statistical approaches is not the amount of human work involved but the nature of the work.[13]

For this project, paraphrases were acquired and organized according to linguistic principles. This organization is deliberately loose since its purpose is only to support explaining the automatically generated paraphrases. We resisted the temptation to create a fine-grained classification to avoid repeating the practice of many language studies where classification is mistaken for theory.

Paraphrasable entities can be:

1. single words in a fixed form: inherently ↔ intrinsically
2. single words that require the part of speech to be checked: $ache_{Noun}$ ↔ $aching_{Noun}$
3. single words that allow for morphological variation, so the word's morphological features must be analyzed and then matched in the output version: *mend → *repair (an asterisk indicates that the word is a verb that can be inflected)
4. multiword expressions consisting of strings that might involve any of the above types of variability: *cause damage → *cause harm
5. multiword expressions with variable slots: *commit to V-infin ↔ *pledge to V-infin
6. expressions that can have different ordering: always [clause-internal or clause-final] ↔ all the time [clause-final only]
7. multiword expressions with variable slots that require coreferences to be checked:
   As far as NP1 is concerned, Pro1 VP ↔ As far as NP1 goes, Pro1 VP
   As far as Harry is concerned, he likes the idea. ↔ As far as Harry goes, he likes the idea.
8. syntactic structures amenable to transformations, such as clauses that can be passivized or unpassivized, and sentences containing main and subordinate clauses whose ordering can be switched. This can require changing the referring expressions in a chain of coreference. For example, the original (10a) cannot be paraphrased by (10b) but it can be paraphrased by (10c).

(10) a. Even though my bag was too heavy, I carried it all the way to the dorm.
b. * I carried it all the way to the dorm even though my bag was too heavy.
c. I carried my bag all the way to the dorm even though it was too heavy.

As we have already explained, there are various ways that paraphrasable entities in our database can be classified, such as by the number of constituents in the construction, the nature of the constituents (e.g., strings vs. variables), the kind and extent of processing required to do the paraphrase (e.g., string replacement vs. syntactic transformation), the syntactic status of the constituent (e.g., noun phrase, verb phrase, adjective), the unidirectionality or bidirectionality of variants, and so on. In this section we present an informal sampling of what our repository contains. By default, verbs can be conjugated and singular nouns can occur in the plural. These details are specified in the repository but omitted here for the sake of readability.

[13] We describe why we think that the kind of manual work we are doing holds much greater promise for AI than the kind pursued in most of today's NLP in McShane and Nirenburg (2021) and McShane, Nirenburg, and English (2024).

Table 2. An informal sampling of paraphrases in our repository.

Label	Example
PROPOSE-PLAN	I think we should VP ↔ I propose that we VP
REQUEST-INFO-YN	Did Subj VP_{Infin}? ↔ Has Subj $VP_{PastPart}$?
OBLIGATIVE (value 1)	Subj has to VP ↔ Subj needs to VP
PHASE (value BEGIN)	Subj is starting to VP ↔ Subj is beginning to VP
passive → active	The movie star was hounded by the press. → The press hounded the movie star.
Subject ellipsis in clausal coordination	We had a nice meal and then **we** talked for a while. → We had a nice meal and then ___ talked for a while.
complementizer ellipsis	[for verbs like *acknowledge, allege, assume, believe, claim, conclude, decide, discover, doubt, expect*] I expect he'll come. ↔ I expect that he'll come.
main-subordinate ordering	Because he was tired, he didn't go. ↔ He didn't go because he was tired.
was vs. *got* passivization	[for verbs including accused, annihilated, banned, acquitted, appointed, blamed, adopted, approved, built, etc.] The building was built fast. ↔ The building got built fast.
people vs. pleonastic *it*	[for verbs including acknowledge, admit, assert, confirm, decide, demonstrate, etc.] People admit that the plan was bad. ↔ It was admitted that the plan was bad.
prep-part ordering	[for phrasal verbs including add up, back up, cover up, carry around, drag around, wave around, and many, many more] He backed up the car. ↔ He backed the car up.
Colloquial informalities	You see, CL → CL
Discourse-smoothing fillers	At the end of the day, CL → CL
plain nouns	abilities ↔ capabilities
plain adjectives	famished ↔ extremely hungry
plain adverbs	again and again → repeatedly
plain verbs	counter ↔ counteract
plain conjunctions	despite the fact that → although
Paraphrases involving negation	this is no place for ↔ this isn't any place for
Paraphrases involving *there is*	There is a danger ↔ There is a risk
Nominal compound vs of-PP	lightning bolt ↔ bolt of lightning
Adj reordering	able and willing ↔ willing and able
the topic of/the issue of/nil	address the issue of NP ↔ address the topic of NP ↔ address NP
Multiword expressions identified through roundtrip translation and not further classified	at high speed ↔ at great speed // serves as a reminder that → reminds us that // the seriousness of ↔ the serious nature of // not as of yet ↔ not yet // assassination → killing // assassination → murder // financial status of ↔ financial condition of // over the past few years ↔ in recent years

At the time of writing, the repository contains 1912 paraphrase sets. It is available upon request for research purposes.

3. The System and Its Assessment

The anonymizer works as follows. (1) It morphologically and semantically analyzes the input text using the spaCy parser (Honnibal & Montani, 2017). (2) It carries out all available string-level

paraphrases. If multiple replacements are available, it generates candidate outputs for all of them, resulting in a new set of sentences. (3) It re-analyzes the new set of sentences using the spaCy parser to account for any changes made in the first phase. (4) It carries out all available variable-inclusive paraphrases on all candidate sentences. (5) It re-analyzes those sentences using spaCy. (6) It runs applicable syntactic transformations on the candidate sentences. This results in a text in which some sentences will not have been touched at all, some will have exactly one output paraphrase, and some will have more than one candidate paraphrase. (7) The system then calls an LLM to prune out any odd sentences. If this results in no paraphrase candidates for a given sentence, the original sentence is reverted to. If this results in more than one paraphrase candidate, then the LLM is asked to select among them.

How well the anonymizer performs is almost wholly dependent upon the size of the paraphrase database. Only *almost* wholly because parsing errors can lead to paraphrasing errors. Since the paraphrase database is currently relatively small, the anonymizer is best described as either (a) a free-standing proof-of-concept system or (b) one module of a nascent multi-pass hybrid system (cf. Section 1). Accordingly, it would not make sense to formally test whether it can independently mask authorship – it likely cannot; and we don't have access to an authorship attribution system to test it. So, although we do provide a sample system run, the main contribution of this work lies elsewhere.

One point of assessment is whether our core hypotheses were correct – and, indeed, they were:

- there *do* exist expressions and constructions in English that are reliable paraphrases for one another in any context, without the need for semantic analysis;
- this approach to paraphrasing *is* explainable;
- a trained linguist *can* acquire a repository of useful paraphrases quickly;
- these paraphrases *can* be operationalized in a paraphrasing system; and
- LLMs *are* useful in weeding out paraphrasing errors and selecting the best from multiple candidate paraphrases.

Another point of assessment is whether we can iteratively improve the system during testing, which we easily can. For example, during an early testing run we identified errors like the ones shown below, which involve fixed expression being treated as compositional (11-13) and the need to lexically and/or syntactically constrain the context of a paraphrase to make it reliable (14-20).

(11) The air national **guard** was stationed… → The air national **watchman** was stationed…
(12) It is a **naked** truth that there was … → It is a **nude** truth that there was …
(13) He … got a **mandatory** life sentence reduced to probation → He … got an **obligatory** life sentence reduced to probation. [5/6; this is still fully interpretable and many readers might not even realize that *mandatory life sentence* is a fixed expression; so, all of the LLMs arguably did fine.]
(14) He is **not known** to be the kind of person to … → He is **unknown** to be the kind of person to …
(15) **Because** of communication difficulties, … → **Given that** of communication difficulties, …
(16) But call them college savings bonds and parents can't buy **enough** of them. → But call them college savings bonds and parents can't buy **sufficient** of them.
(17) The improvements in Medicare are very **real**. → The improvements in Medicare are very **actual**.

(18) You had sexual relations, with a man you were **not married** to. → You had sexual relations, with a man you were **unmarried** to. [5/6]

(19) One of the reasons our children are doing so well is **because** we hold people accountable. → One of the reasons our children are doing so well is **given that** we hold people accountable. [5/6]

(20) But a careful look at the history of sanctions **suggests** that they only succeed when you follow certain guidelines. → But a careful look at the history of sanctions **recommends** that they only succeed when you follow certain guidelines. [3/6]

We tested the same six LLMs mentioned in Footnote 4 to see if they would confirm that these sentences are bad English—and, indeed, most of them did. When not all six detected the problem, the number that *did* detect the problem is indicated as [#/6] following the sentence. The results suggest that we can count on LLMs to do a pretty good job of blocking infelicitous paraphrasing results without our having to spend undue time tweaking the human-compiled paraphrasing knowledge base. We present these examples both to give a taste of the knowledge engineering process (we amended the knowledge and/or processing rules to avoid such errors in the future) and to show that we have an LLM-based solution to dealing with residual problems.

Another way to assess this work is a quantitative evaluation. As a first stab at it, we randomly selected 100 examples, 25 from each of the following classes: "cheap" changes (punctuation, contraction, hyphenation, acronyms), "not-so-cheap" string-level patterns, variable-inclusive constructions, and transformations. The system got either 95/100 or 99/100, depending on whether or not one considers examples (5)-(8), discussed above, to be acceptable paraphrases. The full results are available in an online appendix at https://docs.google.com/spreadsheets/d/10g0rTYjYgfOldjKKC0YSDtG-iBy__DTN_78vXN2kKcY/edit#gid=851271307.

A final point of assessment involves the broader impacts of this R&D effort. First, the paraphrasing database and system are being integrated into the OntoAgent knowledge environment that supports the development of all of our lab's agent systems. Second, the paraphrase database is available upon request to others for research purposes. Third, we are currently working on integrating this knowledge-based anonymizer with an LLM-based anonymizer in the way noted in Section 1. Finally, we believe that our paraphrase database can be used to provide post-hoc explanations of certain aspects of that LLM-based anonymizer. This process works as follows. The LLM-based anonymizer paraphrases text with no reference to our paraphrase database. However, the source and target versions of the text can be compared against our database. If, for a given text span—which may be more than a single sentence—the source and target versions contain different paraphrases included in our database, then that means that the author wanted to express this meaning and the LLM has paraphrased it as part of the anonymization. Naturally, our paraphrase sets will not cover all changes that the LLM will introduce into the text; but for the ones it does, explanations are available, as explained in tables 1 and 2. Of course, these explanations do not address *why* the LLM did what it did – that is not known; but it does give some useful insights into which meanings were paraphrased. The use of post hoc "explanations" of machine learning systems is not new – in fact, it is at the heart of the so-called XAI (explainable AI) movement.[14]

[14] As Babic et al. (Babic et al., 2021) explain, XAI research has concentrated on "*post hoc* algorithmically generated rationales of black-box predictions, which are not necessarily the actual reasons behind those predictions or related causally to them... [and which] are unlikely to contribute to our understanding of [a system's] inner workings."

4. Contributions to Cognitive Systems Research

As we have explained, the reported anonymization system is not a full-blown cognitive system. It can't be because it must operate over unrestricted text, for which full semantic analysis is beyond the state of the art. However, there are three ways in which this work contributes to the field of cognitive systems.

1. This project shows that short-term gains can be achieved within a program of R&D primarily aiming at long-term objectives. That is, it would have been impossible to design and implement the knowledge-based anonymization system we describe here within a small-scale, short-term project without the linguistic and ontological resources of the OntoAgent content-centric cognitive architecture and the decades-long experience of developing semantically-oriented language processing systems. Our anonymizer captures salient features of how paraphrase works for people, and constrains coverage to paraphrase correspondences that people interpret as being reliable in any context.
2. The three-pronged cognitive model of paraphrase that underlies the work – which covers faithful meaning retention, the management of polysemy, and the definition of author style – reflects theoretical and practical advances in the computational linguistic understanding of paraphrase.
3. The plans for integrating our knowledge-based model with processing by LLM is an innovative approach to neurosymbolic architectures that is worth pursuing both theoretically and for its potential practical gains.

Acknowledgments

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 2022-2207220001. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Abbasi, A., & Chen, H. (2008). Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Trans. Inf. Syst.*, 26(2):7:1–7:29.

Bevendorff, J., Potthast, M., Hagen, M., & Stein, B. (2019). Heuristic authorship obfuscation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1098–1108. Florence, Italy, July 28 - August 2, 2019. © 2019 Association for Computational Linguistics.

Bhagat, R., & Hovy, E. (2013). What is a paraphrase? *Computational Linguistics*, *39*(3), 463 – 472. Association for Computational Linguistics.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Tat Lee, Y., Li, Y., Lundberg, S., Nori, H., Palangi, H., Tulio Ribeiro, M., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv. https://arxiv.org/abs/2303.12712

Burrows, S., Potthast, M., & Stein, B. (2012). Paraphrase acquisition via crowdsourcing and machine learning. *ACM Transactions on Intelligent Systems and Technology*, vol. V, no. N, January 2012, Pages 1–22.

Bybee, J. L. (2013). Usage-based theory and exemplar representations of constructions. In T. Hoffmann & G. Trousdale (Eds.), *The Oxford handbook of construction grammar* (pp. 49–69). Oxford University Press.

Davies, M. (2008-). The Corpus of Contemporary American English (COCA): One billion words, 1990–2019. https://www.english-corpora.org/coca/

Dolan, W. B., & Brockett, C. (2005). Automatically Constructing a Corpus of Sentential Paraphrases. In Proceedings of the Third International Workshop on Paraphrasing (IWP2005), https://aclanthology.org/I05-5002

Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. https://spacy.io/

Mahowald, K., Ivanova, A. A., Blank, I. A, Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). Dissociating language and thought in large language models: A cognitive perspective. arXiv, 2301.06627.

Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics*, 33(3): 299–320. Oxford University Press. doi:10.1093/applin/ams010

McShane, M., Nirenburg, S., and English, J. (2024, *forthcoming*). *Agents in the Long Game of AI: Computational cognitive modeling for trustworthy, hybrid AI.* MIT Press.

Nirenburg, S., McShane, M., & English, J. (2020). Content-centric computational cognitive modeling. *Proceedings of the Seventh Annual Conference on Advances in Cognitive Systems*.