A Neurosymbolic Approach to Authorship Anonymization

Abstract

We report a neurosymbolic approach to authorship anonymization that combines knowledge-based paraphrasing, grounded in cognitive modeling, with support functions provided by a large language model (LLM). The cognitive model accounts for four things: 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, how to define and characterize an author's *style*, and how to leverage *human linguistic capabilities* when preparing systems to automatically anonymize texts. LLMs augment the knowledge-based paraphrases in three ways: by filtering out atypical formulations, by selecting the best from multiple candidate paraphrases, and by offering additional paraphrases in case the knowledge-based paraphrasing fails to adequately anonymize the text. This neurosymbolic architecture favors knowledge-based processing for being reliable and explainable, while exploiting LLMs for what they do best: manipulate regularities in the surface form of language.

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 cannot take this approach because full semantic analysis of open text is beyond the state of the art. This leaves three options.

Option 1: Use machine learning (ML). One can sidestep the need to compute meaning by using ML to paraphrase (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 larger chunks of text.[2] Paraphrasing larger chunks of text can result in texts that are quite different from the original, thus fulfilling the goal of anonymization. However, paraphrasing by LLMs—like all processing 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

---

[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 was pursued by other developers working on the same research project; see the Acknowledgments.

paraphrase "He was disliked by his subordinates because he was too hasty". This is not a felicitous paraphrase on 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]

Option 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 semantic analysis, and then implement a system to carry out those replacements. The cognitive model underlying this approach must account for four things: what it means to faithfully retain *meaning and discourse coherence* in a paraphrase (section 3.1); how to deal with lexical *polysemy* given that the system cannot rely on a full semantic analysis of the text (section 3.2); how to define and characterize an author's *style*, which should be markedly different in the source and paraphrased versions of the text (section 3.3); and how best to leverage *human linguistic capabilities* to prepare systems to automatically anonymize texts (section 3.4). The downside of the knowledge-based 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.

Option 3: Use a neurosymbolic approach. One can combine knowledge-based and LLM-based capabilities into a neurosymbolic system that optimally balances reliability, explainability, and coverage. We believe that this approach will ultimately be the most successful, and continue our story with a brief introduction to just such an architecture.

2. The Neurosymbolic Architecture

Figure 1 illustrates our vision of a neurosymbolic architecture that could optimize automatic authorship anonymization. We present it from the outset since it is the big picture that will contextualize the upcoming discussions of theory and linguistic phenomena.

---

[3] There is precedent for defining *paraphrase* loosely. For example, the Microsoft Research Paraphrase Corpus 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".
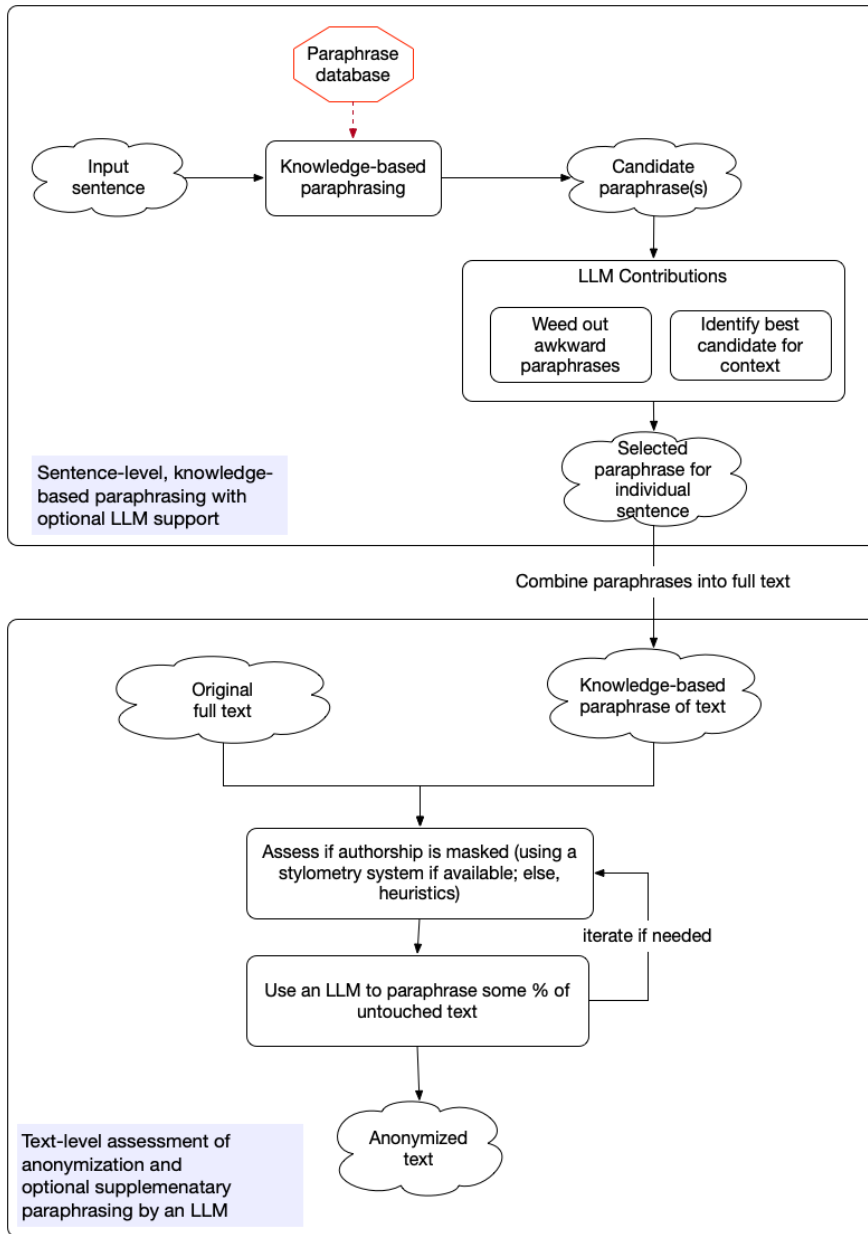
Figure 1. A neurosymbolic architecture for authorship anonymization.

The main processing flow for an input text is as follows:

1. A knowledge-based system paraphrases the text sentence by sentence. This ensures strict paraphrasing, explainability, and quite high reliability, albeit currently limited coverage.

When mistakes occur, they are minor, not the sorts of hallucinations that LLMs can generate. This process can return any number of paraphrases for each sentence, including zero.

2. An LLM vets all paraphrases and weeds out anything that is atypical, which might be due to a mistake in the paraphrasing knowledge base, a mistake in language processing (morphological analysis or generation; syntactic analysis or generation; or the application of paraphrasing rules), or a legitimately unexpected kind of input (such as a proper name that is written without capital letters). Detecting atypical formulations is exactly the kind of thing that LLMs are well suited for since it relies on the statistical likelihood of sequences of strings (Mahowald et al., 2023; Bubeck et al., 2023). For example, at an early stage of this work our paraphrase inventory erroneously included *very* and *really* as bidirectionally replaceable, leading to the erroneous paraphrasing of *Do you really believe…* as *Do you very believe…* An LLM readily detected the problem, allowing the neurosymbolic system to reject this paraphrase.

3. An LLM selects among multiple candidate paraphrases.[4] For example, the knowledge-based paraphraser paraphrased example (1) in three ways, and several LLMs we tested selected "b" as the best.

    (1) **Drivers are not required to** record traffic hazards, though, and apparently **seldom** do so.
    a. **Drivers are not bound to** record traffic hazards, though, and apparently **hardly ever** do so.
    b. **Drivers do not need to** record traffic hazards, though, and apparently **hardly ever** do so. ✓
    c. **Drivers have no need to** record traffic hazards, though, and apparently **hardly ever** do so.

    The LLMs we 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." Note, however, that the LLM's preference should not necessarily always be selected since it might be the one that least modifies the original text. So, a knowledge-based, post-LLM selection process would best introduce some variability into the ultimate selection.

4. All of the selected paraphrases are combined into a new text, which is compared against the original text to determine if authorship has been masked. Ideally, this will be done using a stylometry system but, absent that, heuristics can be used to estimate how different the original and paraphrased versions are.

5. If authorship has been reliably masked, processing ends: the exclusively knowledge-based paraphrase will be highly reliable and fully explainable, which is ideal.

6. If authorship has not been reliably masked, an LLM paraphrases a small percentage of as-yet untouched sentences, either individually or as multi-sentence blocks. This will potentially introduce errors, so the amount of paraphrasing by the LLM should be held to a minimum.

7. Anonymization is reassessed. If authorship is not yet masked, then the LLM paraphrases additional small segments of text until confident anonymization has been achieved.

---

[4] We use an LLM in a similar way during text generation in our cognitive agent system (McShane, Nirenburg, & English, 2024).

We believe that the neurosymbolic architecture in Figure 1 is a promising way to fundamentally solve the problem of authorship anonymization over time. As the database of paraphrases underlying the knowledge-based system grows, the latter will become increasingly able to independently anonymize texts, with a decreasing need for paraphrasing supplementation by LLMs. This is optimal since the knowledge-based paraphrases are more reliable and explainable than what an LLM can generate. To date, we have made significant inroads into operationalizing this vision, as the remainder of this paper details.

One might wonder—as did an audience member at a talk about this work—why not first paraphrase text using an LLM and then vet the paraphrase using knowledge-based methods? This question reflects typical assumptions about neurosymbolic systems, which casts them as primarily stochastic with light supplementation by knowledge-based rules. The problem with an LLM-first approach to paraphrasing is that the knowledge-based clean-up rules would have to account for an open-ended spectrum of mistakes, omissions, additions, and hallucinations that the LLM could potentially introduce. Even if one could classify all such eventualities (in fact, one couldn't), it would be extremely difficult to implement the reasoning to automatically detect them. By contrast, when we use an LLM to check the output of the knowledge-based paraphraser, we are only asking it to assess whether the sentence is typical at the surface level, which is an immeasurably simpler task that LLMs are inherently well-suited to carry out. As concerns our use of an LLM for supplementary paraphrasing, that is a stopgap until the paraphrasing database becomes large enough to support fully knowledge-based paraphrasing. If the LLM introduces a mistake during its supplementary paraphrasing, that is just the cost of producing a system under the fast timeline that is expected of intelligent systems these days.

3. The Four Pillars of the Cognitive Model

As mentioned earlier, the cognitive model underlying this approach must account for four things: what it means to faithfully retain *meaning and discourse coherence* in a paraphrase; how to deal with lexical *polysemy* given that the system cannot rely on a full semantic analysis of the text; how to define and characterize an author's *style*, which should be markedly different in the source and paraphrased versions of the text; and how best to leverage *human linguistic capabilities* to prepare systems to automatically anonymize texts. These are discussed in turn in the subsections below.

3.1 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, context-appropriate English. (2) – (5) are examples of true paraphrases under this definition.

(2) a. **To** do well, you **have** to study hard.   b. **In order to** do well, you **need** to study hard.

(3) a. **Apart from** him, nobody else came.   b. Nobody else came **except for** him.

(4) a. This **led to** a big debate.   b. This **resulted in** a big debate.

(5) 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 any of 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*.[4]

The need to retain discourse coherence means that syntactic transformations cannot be randomly applied.[5] For example, *Charlotte fixed the fence* should not be subjected to the following transformations unless warranted by the larger discourse:

passivization:	The fence was fixed by Charlotte.
subject dislocation:	Charlotte, she fixed the fence.
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 of 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 were not warranted.

Although it is important to not alter the meaning of the original text, one 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 (which must take

---

[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).

[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*).

into consideration that automatic paraphrasing will naturally be error-prone) might be considered acceptable or not acceptable.

(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 was switched, the sequence of coreferential expressions became suboptimal.]
(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 were to find those same choices in a native-language text, you wouldn't think twice.

The paraphrases above would make the author's style less academic, which might be a valuable obfuscation strategy. To operationalize "less academic style", one could, for example, introduce rules to disrupt the 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.

3.2 Dealing with Polysemy In the Absence of Full Semantic Analysis

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 in the absence of full semantic analysis is to focus on the construction-based nature of human languages. That is, a 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.[6]

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

---

[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 lists of multiword expressions.

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 (*Additionally,*), 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.[7]

- 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 the expressions *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 *issue, study, question, 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, development, dimension, challenge, limitation, blow, distinction, refinement, argument, consequence,* or *delay*.
  - *Look$_V$ to Pronoun for* can be paraphrased by *consult$_V$ 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:

---

[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 decreases. 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…" (Davies, 2008-).

- Currently, the word lists associated with constructions are incomplete, having been compiled through a combination of introspection and online search of the COCA corpus (Davies, 2008-). These lists can be expanded given more time. But such word lists are essential because *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 acquiring an ontological-semantic lexicon of the type that our research group is developing for cognitive agent systems. However, for this short-term anonymization project, listing, albeit incomplete, is the only feasible option.
- 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, not every single text component needs to be paraphrased—only enough to mask the author.[8]
- The choice of what to consider a variable versus a constant can be tricky but, for the current purposes, it is based on a fast human judgment. For example, *It's a further* [*reason, example,* etc.] ↔ *It's another* [*reason, example,* etc.] treats *It's* as a constant. There is a separate construction for *That's a further* [*reason, example,* etc.] ↔ *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.[9] For example, given the input *Not too long ago he 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 *he changed jobs* can be paraphrased by *he switched jobs*, *he got a new job,* and *he left his old job for a new one*; and
- a comma after the sentence-initial adverbial is optional.

All of these paraphrase opportunities create a substantial set of strict paraphrases from which an anonymization system could select.

Paraphrases can be reliable in one direction but not the other, which 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 (not every police officer is a woman). 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 the directionality of confident paraphrases must be made by people.

---

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

[10] This could be tested using psycholinguistic experimentation. A relevant direction of research involves how construction frequency interacts with memory (e.g., Bybee, 2013: p. 49).

Finally, there are standard ways of saying things, and switching out components of a canonical expression can result in an unnatural formulation. For example, replacing *I would appreciate it if you would…* by *I would value it if you would*… sounds hyper-formal, even though it is grammatical and understandable. Similarly, although changing the ordering of adjectives can lead to a meaning-preserving modification of an input, 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.

3.3 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.[10] 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 put it another way:

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, **a writer has a choice when expressing this meaning**.
3. A writer's **preference** for how to express this meaning is **a stylistic feature.**
4. Every time a writer uses an expression in our database, that **choice** involves not choosing the other option(s).
5. The **inventory of choices** for expressing the meanings in the paraphrase database are the **writer's profile** – or, more specifically, the aspect of the writer'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 the paraphrase 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 OntoAgent ontology, which is among the core knowledge bases used by our

---

[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) relies on an unnecessarily narrow definition of style.

research group's cognitive agent systems: e.g., EXPRESS-EMPHASIS (see section 5, point 5, for details). And in still other cases, a proxy label is used that will suffice until such time as we expand the OntoAgent 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.[11]

Feature	Sample values (i.e., paraphrase sets)
Conventional linguistic functions	
Active/passive	Subj V DirectObj ↔ Subj$_{UnderlyingDirectObj}$ *be* V$_{PastPart}$ by NP$_{UnderlyingSubj}$
Overt or elided subject in coordinated clauses	Subj$_1$ CL$_1$ and (Adv) Pro$_1$ CL$_2$ ↔ Subj$_1$ CL$_1$ and (Adv) ___ CL$_2$
The ordering of particles that are homographous with prepositions	[for non-pronominal DirectObjs only] • carry around DirectObj ↔ carry DirectObj around • drag around DirectObj ↔ drag DirectObj around
Nominal compound vs. prepositional phrase with 'of'	• gas shortage ↔ shortage of gas • depression risk ↔ risk of depression
The ordering of conjoined adjectives	• bright and lively ↔ lively and bright • calm and smooth ↔ smooth and calm
The presence or absence of 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]	inherently ↔ intrinsically
ProxyNoun:[acquisition of, acquiring of]	acquisition of ↔ acquiring of
ProxySubjV:[this involves, this entails]	this involves ↔ this entails
ProxyV:[affects, has an effect on]	affects ↔ has an effect on
ProxyAdj:[thorough, extensive]	thorough ↔ extensive

---

[12] Typical linguistic abbreviations are used: V (verb), VP (verb phrase), CL (clause), Adj (adjective), Adv (adverb), N (noun), Subj (subject), DirectObj (direct object), PastPart (past participle).

3.4 Leveraging Human Linguistic Capabilities to Build Anonymization Systems

The final pillar of our cognitive model of paraphrasing involves determining how best to use people to develop the paraphrase database. This means leveraging their knowledge about ambiguity and paraphrasing while offering them engaging work that does not impose an overly heavy cognitive load. Naturally, automation must be used in all ways possible.

As a reminder, the goal is to create a large inventory of words, phrases, open patterns, and syntactic transformations that are interchangeable in all contexts without the need for semantic analysis. The replacements must be specified as either bidirectional (*maybe ↔ perhaps*) or unidirectional (*nation → country*). We primarily used four knowledge-acquisition methodologies, which we briefly describe in turn.

Method 1. We used the OntoAgent lexicon as a source of ready-made paraphrases and the OntoAgent ontology to guide the search for others. To give just a few examples:

- The concept PROPOSE-PLAN can be expressed as *I think we should VP, I propose that we VP, I think it would be a good idea to VP*, etc.
- Obligative modality with a value of 1 (on the scale {0,1}) can be expressed as *Subj has to VP, Subj needs to VP, Subj is obliged to VP*, etc.
- The starting phase of an event can be expressed as *Subj is starting to VP, Subj is beginning to VP, Subj has just started to VP,* etc.
- Complementizer ellipsis is permitted for some verbs, including *acknowledge, allege, assume,* and many more: *I assume (that) she left on time*.
- Some subordinating conjunctions, like *because*, permit either clause order: *Because he was tired, he didn't go. ↔ He didn't go because he was tired.* These alternations are not pragmatically perfect in all cases but we have to balance the need to obfuscate with the desire for the text to sound as good as possible.
- Some verbs—such as *accused, annihilated, banned, built,* etc. —permit passivization using *got* in addition to passivization using *be*: *The building was built fast. ↔ The building got built fast.*

Method 2. We found lists of frequent adverbs, time expressions, phrasal verbs, etc., on the internet and then thought up or looked up (using online thesauri) paraphrases for them. It is important to focus on frequent expressions in order to have a reasonable chance of having hits in texts that need to be anonymized.

Method 3. We did roundtrip machine translation of texts—English → French → English—in batch mode using the online translation tool called DeepL (deepl.com). We then used the online Diffchecker text comparison tool (https://www.diffchecker.com/text-compare/) to highlight differences between the original and final English versions. Finally, we manually scanned the highlighted segments for useful paraphrases. Although not a large portion of the highlighted

correspondences could be directly included in our paraphrase database, this was a useful method of jogging acquirers' memories for paraphrases that were, in fact, useful.

Method 4. We used the online search engine for the COCA corpus (Davies, 2008-; https://www.english-corpora.org/coca/) to test and expand upon linguistic hypotheses about classes of paraphrasable entities. For example, many adverbs in "ly" can be paraphrased using the constructions *in a ADJ manner, in a ADJ way* and *in a ADJ fashion*. This offers four-way paraphrase sets like:

differently ↔ in a different manner ↔ in a different fashion ↔ in a different way
consistently↔ in a consistent manner ↔ in a consistent fashion ↔ in a consistent way

The question was, which other adjective-adverb pairs work this way? The COCA search tool helped us to answer that question by providing the output shown in Figure 2, which is a small subset of expressions returned by the query "in a ADJ manner".

Figure 2. A subset of results returned from searching for the pattern "in a ADJ manner" using the online search engine for the COCA corpus (https://www.english-corpora.org/coca/; Davies, 2008-).

We then scanned that list, mentally checking whether all four versions were available for the given adjective in the intended meaning. In some cases they were not: for example, the adverbs *civilizedly, friendlily*, and *certainly* don't work (the latter has a different meaning). However, even for these, the three-way *manner fashion/way* versions are useful paraphrases. What is interesting about the human language faculty is that one can make such judgments in a split second with minimal cognitive load. What takes more knowledge and experience is coming up with fruitful linguistic hypotheses to begin with.

It is not possible to automatically detect context-independent paraphrase equivalents using machine learning, data analytics, or LLMs alone. And, even if one were to try to use those methods, it would require no small amount of manual labor in the form of text annotation, cleaning datasets,

prompt engineering, and so on. So, what differs between knowledge-based and empirical approaches is not the amount of human work involved but the nature of the work.[12]

For this project, we acquired and organized paraphrases according to linguistic principles but rather loosely, with three practical goals in mind: (a) making acquisition efficient, (b) organizing the database according to the processing needed by the paraphrasing engine, and (c) preparing the system to automatically explain the paraphrases. We resisted the temptation to create a fine-grained classification in order to avoid repeating the practice of language studies in which classification is mistaken for theory.

At the structural level, 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 generated 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:
   (10) a. As far as NP_1 is concerned, Pro_1 VP   ↔ As far as NP_1 goes, Pro_1 VP
         b. 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 unpassivized, and sentences containing main and subordinate clauses whose ordering can be switched. Transformations can require changing the referring expressions in a chain of coreference. For example, (11) cannot be paraphrased by (11a) but it can be paraphrased by (11b).

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

As we have shown, 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 natural language processing 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	remember → recall
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
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 – i.e., meanings for which more than one paraphrase is listed.

4. Sentence-Level Paraphrasing and Its Assessment

The sentence-level, knowledge-based paraphraser (top of Figure 1) works as follows:

1. It morphologically and syntactically 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 pass.
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 any number of paraphrases, including zero. How well the knowledge-based paraphraser performs is almost wholly dependent upon the size of the paraphrase database. Only *almost* wholly because errors in morphological or syntactic analysis can lead to paraphrasing mistakes, as can the application of syntactic transformations on types of input that our rules did not foresee (language is anything but fully anticipatable!).

Since the paraphrase database is currently relatively small, it is premature to formally test whether the knowledge-based system can independently mask authorship. It likely cannot; and, unfortunately, we do not have access to an authorship attribution (stylometry) system to test that out. However, we did do a small evaluation focusing on accuracy. Specifically, we randomly selected 100 sentences from an open corpus for which the system could provide paraphrases, 25 from each of the following classes:

1. cheap changes, which are changes in punctuation, contraction, hyphenation, and acronyms
2. not-so-cheap, string-level patterns: again and again → repeatedly
3. variable-inclusive constructions: At the end of the day, CL → CL (i.e., *At the end of the day* is removed)
4. transformations: Because he was tired, he didn't go. ↔ He didn't go because he was tired.

The system got either 95/100 or 99/100, depending on how strictly one judges the quality of the paraphrased text. For example, consider the following paraphrase, in which a single instance of potential modality (*might*) is replaced by a doubling up of potential modality (*there is a possibility that ... might*). The replacement is correct except for the tense: *is* should be *was* in order to preserve the necessary sequence of tenses.

(12)   a.   I knew that that **might** make her more likely to speak, and I still did it because I had to limit the contact.

         b.   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.

Once we detected this mistake, we fixed the rule so that sequences of tenses are now accounted for. But during any evaluation, there are likely to be such *close but not quite* cases. The full results of

the 100 sentence evaluation run are available in an online appendix at https://leia-lab.github.io/ACS2024-Author/.

Another important point of assessment is whether the system can be iteratively improved. The answer is an emphatic *yes*, which is one of the advantages of knowledge-based approaches over empirical ones. For example, during an early test run we identified errors like the ones shown below. Some involve a fixed expression being treated as compositional (13-15), and others require lexical or syntactic constraints on the expression in order to ensure reliable paraphrasing (16-22). We tested the same six LLMs mentioned earlier (Claude 3 Opus, Claude 3 Sonnet, Claude 3 Haiku, GPT-4, GPT-3.5, and Gemini Advanced) to see if they would confirm that these sentences were 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.

(13)   The air national **guard** was stationed… → The air national **watchman** was stationed…

(14)   It is a **naked** truth that there was … → It is a **nude** truth that there was …

(15)   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 some readers might not even realize that *mandatory life sentence* is a fixed expression; so, all of the LLMs arguably did fine.]

(16)   He is **not known** to be the kind of person to … → He is **unknown** to be the kind of person to …

(17)   **Because** of communication difficulties, … → **Given that** of communication difficulties, …

(18)   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.

(19)   The improvements in Medicare are very **real**. → The improvements in Medicare are very **actual**.

(20)   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]

(21)   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]

(22)   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]

These and other results of system testing 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 *did* amend 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 errors in the paraphrasing database.

5. Results

The cognitively-grounded, neurosymbolic model of automatic authorship anonymization presented here was developed under funding that was discontinued midway through the implementation of the vision presented in Figure 1. We implemented the knowledge-based paraphrasing shown in the upper box of Figure 1 but did not advance to integrating those results with LLM-based paraphrasing loop in the lower box. Nevertheless, this R&D effort produced noteworthy contributions on the theoretical, linguistic, and system-building fronts.

1. **We developed a neurosymbolic approach to authorship anonymization** that integrates the best of what knowledge-based methods and LLMs have to offer and charts a path toward achieving ever more reliable and explainable anonymization results over time.

2. **We developed the cognitive model underlying the knowledge-based paraphrasing system.** This model centrally addresses 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, how to define and characterize an author's *style*, and how to leverage *human linguistic capabilities* when preparing systems to automatically anonymize texts.

3. **We fleshed out islands of non-ambiguity in English.** Perhaps *the* biggest challenge in computational semantics is ambiguity, and our group has spent decades working on this problem. However, prior to this project, we had not specially thought about creating an inventory of words, phrases, and constructions that are *not* ambiguous, and using them as islands of confidence to support automatic ambiguity resolution in various language processing subtasks.

4. **We established that automatic knowledge-based paraphrasing *does* work** since language *does* offer unambiguous constructions that can be reliably replaced by other constructions without the need for semantic analysis.

5. **We are incorporating our new paraphrasing system into our research group's main line of work: building Language-Endowed Intelligent Agents (LEIAs) within the HARMONIC cognitive architecture** (Nirenburg, McShane, & English, 2020; McShane, Nirenburg, & English, 2024; Oruganti, 2024). LEIAs understand and generate text using deep semantics and pragmatics, which is beyond the state of the art for open text, thus explaining why we did not use LEIAs for the paraphrasing project. However, the new paraphrasing capabilities we developed are helping us to reducing combinatorial explosion during language generation. To understand how, one needs a bit of background.

   For LEIAs, text generation starts from generation meaning representations (GMRs), which are ontologically-grounded (concept-based) representations of what the agent wants to say. For many meanings, the agent's lexicon contains multiple lexical realizations, which can lead to the generation of many candidate sentences to express a given meaning. To reduce complexity, when the agent is initially composing a sentence out of all of the meanings in the GMR, it uses a default word or phrase to express each meaning. For example, for PROPOSE-PLAN the default is "I think we should …", resulting in sentences like "I think we should replace the battery." Once the generation system has created *a* sentence to convey the meaning of the whole GMR, it can optionally create paraphrases using our new paraphrase generator. For example, "I think we should replace the battery" can be expanded into "I think it would be a good idea to replace the battery", "I propose we replace the battery", "I think it would make sense to replace the battery", and so on. Finally, we ask an LLM to determine which of the paraphrases sounds best in the given context. (For details on our NLG system, see McShane, Nirenburg, and English, 2024, section 4.3.) Building useful agent systems

requires thinking about things like when and how to handle complexity. By contrast, theoretical work can float above such cares but might never transition to practical systems.

6. **We showed that a long-term program of R&D can offer near-term utility.** 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 (a) the linguistic and ontological resources of the LEIA content-centric cognitive architecture, (b) our decades-long experience of developing semantically-oriented language processing systems, and (c) our recent experience in building neurosymbolic architectures (Oruganti et al., 2024). Although this is not the place to delve into linguistic theory, the human-inspired computational theory of semantics called Ontological Semantics (Nirenburg & Raskin, 2004), which underlies LEIA language processing, drove how we conceptualize the mapping between syntax and semantics, which is at the core of the reported approach to paraphrasing.[13]

7. **We demonstrated reliable uses of LLMs within a novel neurosymbolic architecture.** In the knowledge-based anonymization process, LLMs are incorporated (a) to make sure that paraphrases generated by the knowledge-based system sound normal, and (b) to select the best from multiple candidate paraphrases. These are exactly the kinds of things that LLMs are best suited to do since they involve judgments based on word frequencies. By contrast, having LLMs supply supplementary paraphrases (in the lower part of Figure 1) goes beyond their zone of reliable competence; but this is only a stopgap, aimed at configuring a useful system fast. Under our vision, as the knowledge-based component achieves better coverage over time, the need for paraphrasing by the LLM will be phased out and the confidence and explainability of the overall system will increase.

8. **We compiled a database of paraphrase correspondences that is available for research purposes** (contact us).

6. Conclusions

In an AI environment dominated by short-term engineering approaches to individual ("silo") problems, we must not lose sight of opportunities to investigate whether problems lend themselves to scientific treatment and whether efforts to solve them have the potential to make broader contributions. For the problem of text anonymization, we have shown that the answer is a resounding yes.

Even within a purely LLM-based approach to text anonymization, our paraphrase database will be useful in providing post-hoc explanations of certain anonymization results by comparing the source and target versions of the text against our database. If, for a given text span—which might be more than a single sentence—the source and target versions contained paraphrases attested in

---

[14] There are also non-computational theories that address the syntax-to-semantics mapping, which have arisen since the formulation of Ontological Semantics and are largely compatible with it, such as the various flavors of Construction Grammar (Hoffmann & Trousdale, 2013) and Jackendoff's (2023) Parallel Architecture.

our database, that would mean that the author wanted to express the given meaning and that the LLM had adequately paraphrased it as part of the anonymization process. Naturally, our paraphrase sets will not cover all changes that the LLM will introduce into the text; but for the ones it does, explanations will be available, as illustrated by tables 1 and 2. Of course, these explanations do not address *why* the LLM did what it did—that is not known; but they could give some useful insights into which meanings were paraphrased. Using *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]

We hope that this work will spur developers to think about seeking solutions to many kinds of language processing problems that do not wholly or primarily rely on LLMs because, despite the latter's impressive performance on some tasks, their lack of reliability and explainability presents serious challenges to people who are ultimately responsible for the performance of automated systems in high-risk domains.

Acknowledgments

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.

Babic, B., S. Gerke, T. Evgeniu and I.G. Cohen. (2021). Beware explanations from AI in health care. *Science*, 373:6552.

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.

---

[15] As 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."

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

Hoffmann, T., & Trousdale, G. (Eds.). (2013). *The Oxford handbook of construction grammar*. Oxford University Press.

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

Jackendoff, R. (2023), The Parallel Architecture in Language and Elsewhere, Topics in Cognitive Science, Wiley. https://doi.org/10.1111/tops.12698

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). *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*.

Nirenburg, S., & Raskin, V. (2004). *Ontological Semantics*. MIT Press.

Oruganti, S., Nirenburg, S., McShane, M., English, J., Roberts, M. K., & Arndt, C. (2024). HARMONIC: Cognitive and control collaboration in human-robotic teams. *arXiv preprint arXiv:2409.18047*.