

Semantic Annotation for Mapping Abstract Concepts: A Category-Based Approach to Adolescent Well-Being in Contemporary American Young Adult Literature

Ayako Yamashiro

Okinawa Prefectural College of Nursing
1-24-1 Yogi, Naha, Okinawa 9028513 JAPAN
ayamashiro@okinawa-nurs.ac.jp

Abstract

This study develops and internally audits a theory-led semantic annotation procedure for mapping a multidimensional construct to corpus-retrievable semantic tags. Ross et al.'s (2020) five-domain framework of adolescent well-being serves as the test case, with the UCREL Semantic Analysis System (USAS) providing the intermediate representation. Content words from the published domain definitions were annotated and assigned to each originating well-being domain, preserving multi-domain membership. A separate synonym-derived inventory, based on screened thesaurus synonyms, served as a sensitivity condition. Three set-based audits examined containment and novelty, intended versus cross-domain alignment, and overlap among definition-derived domain sets. After reconciliation, deduplication, and exclusion of discourse-marker and unmatched outputs, the definition-derived inventory contained 128 unique semantic tags and the synonym-derived inventory contained 63. A total of 44 tags were shared and 19 were novel (containment = .698; novelty = .302; Jaccard = .299). Intended-domain alignment was strongest for 3 of 5 domains, and definition-derived overlap ranged from .210 to .328. The findings support the definition-derived inventory as the primary, traceable mapping and the synonym-derived inventory as a separate sensitivity condition. The procedure identifies transparent semantic indicators of the theoretical construct rather than measuring adolescent well-being directly.

Keywords: adolescent well-being, corpus linguistics, semantic annotation, mapping audit, Jaccard similarity

1. Introduction

Contemporary American young adult literature represents adolescence through embodied experience, relationships, institutional conditions, learning, identity, agency, and adversity. These dimensions make adolescent well-being an analytically productive organizing concept, but they also expose a measurement problem. Well-being is a multidimensional theoretical construct that cannot be encapsulated by a single term. Focusing solely on well-being would emphasize explicit mentions while overlooking the broader semantic expression of the concept. Conversely, using an unrestricted dictionary would lead to a wider range of retrievals, but it would blur the distinction between adolescent well-being and general evaluative or thematic language.

The conceptual literature reinforces the need for a multidimensional approach. Langford et al. (2021) propose a US-focused ecological framework for youth and young adults comprising physical health and safety, cognitive and mental health, social and emotional well-being, economic well-being, and racial and ethnic equity. Their model places the young person within family, community, private-sector, and public-system environments and emphasizes well-becoming across the transition to adulthood (pp. 17–18). Ross et al. (2020), drawing on literature review and consultation across international agencies and youth networks, define adolescent well-being through five interconnected domains. Their expanded definition states that “adolescents have the support, confidence, and resources to thrive in contexts of secure and healthy relationships, realizing their full potential and rights” (p. 473).

Both frameworks reject a narrow equation of well-being with affect or health alone, but they serve different analytic purposes. Langford et al. (2021) foreground the ecology of

youth and young adults as well as structural equity; Ross et al. (2020) provide an adolescent-specific, internationally developed framework whose domain titles, subdomains, requirements, and well-being types supply an explicit textual basis for annotation. The present mapping therefore uses Ross et al. (2020) as its source taxonomy while retaining Langford et al. (2021) as evidence that holistic, contextualized well-being is also salient in the youth and young adults context.

Corpus annotation offers a method for moving from a conceptual framework to inspectable textual evidence. Corpus linguistics can be understood as a set of methods for studying language through machine-readable corpora; computational tools enable rapid and reliable searching, frequency profiling, and concordancing, but the evidence they retrieve still requires qualitative linguistic interpretation (McEnery & Hardie, 2012, pp. 1–2; Rayson, 2008, p. 528). The UCREL Semantic Analysis System (USAS) supplies a hierarchical semantic tagset that groups related word senses into semantic fields, including lexical items and multiword units (Archer et al., 2002, pp. 1–2). Wmatrix provides an interface for automatic part-of-speech and semantic annotation, frequency profiles, keyness comparisons, and concordances (Rayson, 2008, p. 529). Rayson's later poster describes the extension of the key semantic domains method in Wmatrix beyond English corpora (Rayson, 2021, Abstract section; Key Semantic Domains section). At this intermediate semantic level, lexical items and multiword expressions that may be individually low in frequency can be grouped under shared semantic-domain labels, making broader patterns more visible than a surface word list alone while preserving access to the lexical items and concordance lines through which each category is instantiated (Rayson, 2008, pp. 542–543).

The semantic annotation system, however, is general rather than a domain-specific measure of adolescent well-being.

Its semantic categories group word senses at a level of semantic generality, and its sense distinctions are sometimes coarser than those required by a specialized study (Archer et al., 2002, pp. 1–2; Rayson, 2008, p. 529). A crosswalk from theory to semantic tags should therefore be treated as an operational mapping, not as an equivalence between taxonomies. This caution is especially important for literary data. Corpus analysis can disclose textual features that a researcher may not otherwise consciously notice, but textual findings do not directly determine literary effects or literary significance (Widdowson, 2008, p. 293). In this study, therefore, a semantic tag should be treated as an indicator of candidate textual evidence.

The central methodological task is consequently validation of the mapping procedure itself. Automated text methods amplify human analysis but remain dependent on problem-specific assumptions and checks (Grimmer & Stewart, 2013, pp. 2–4). Dictionaries may misclassify language when transferred beyond the domain for which they were developed (Loughran & McDonald, 2011, pp. 35–36), while preprocessing choices can materially alter downstream findings (Denny & Spirling, 2018, pp. 1–4). For the present mapping, transparency requires a clear separation between semantic tags anchored directly in Ross et al.’s published framework and tags introduced through synonym expansion. It also requires an audit that does not overstate descriptive set overlap as inferential evidence of validity.

This study suggests and internally audits a replicable mapping for retrieving candidate semantic evidence related to Ross et al.’s (2020) five domains from contemporary American young adult literature. The audit precedes any claim about the frequency, distribution, or literary function of those domains in a corpus. It addresses the following research questions:

- RQ1. To what extent are semantic tags in the synonym-derived inventory contained within the definition-derived tag inventory, and how much novelty do they introduce?
- RQ2. Does each domain’s synonym-derived tagset align most strongly with the definition-derived set for its intended domain rather than with another domain?
- RQ3. How much overlap exists among the five definition-derived domain sets themselves?

2. Conceptual and Semantic Basis

2.1 Ross et al.’s Five Domain Framework

Ross et al.’s (2020) framework comprises D1, good health and optimum nutrition; D2, connectedness, positive values, and contribution to society; D3, safety and a supportive environment; D4, learning, competence, education, skills, and employability; and D5, agency and resilience. The framework is suitable for semantic operationalization for the following reasons. They encompass both subjective and objective dimensions of well-being and are underpinned by considerations of gender, equity, and rights (p. 473). It further specifies subdomains, conditions or requirements, and physical, nutritional, emotional, cognitive, and sociocultural types of well-being (pp. 474–475). This layered structure permits a definition-derived lexical template that is broader than the five domain titles yet remains anchored in published material.

The resulting categories should not be treated as direct observations of a latent psychological state. A fictional passage may invoke safety, education, identity, or relationships without asserting that a character possesses positive well-being. It may portray provision, deprivation, conflict, irony, memory, aspiration, or another character’s perspective. Accordingly, the target of annotation is textual evidence semantically related to the framework, not well-being itself.

2.2 Semantic Annotation as an Intermediate Representation

The annotation system groups word senses connected at a general conceptual level. Its lexical groupings may include synonyms, antonyms, hypernyms, and hyponyms, and its lexicon associates a syntactic tag with one or more semantic tags (Archer et al., 2002, pp. 1–2). The guide describes 21 major discourse fields subdivided into 232 category labels, with additional markers for semantic polarity, degree, multiword units, and multiple membership (Rayson et al., 2004, p. 3). Wmatrix combines semantic annotation with part-of-speech tagging, frequency lists, concordances, and corpus comparison (Rayson, 2008, pp. 527–529). This annotation framework is useful here because it replaces a brittle surface-word list with a documented semantic level that can be inspected through its lexical realizations.

The relationship between the semantic tagset and the well-being framework is many-to-many. A single Ross domain can invoke numerous semantic tags, and a semantic tag can be relevant to more than one Ross domain. For example, social affiliation, power, evaluation, action, and psychological processes may recur across connectedness, safety, learning, and agency. Forced assignment of every semantic tag to only one well-being domain would therefore impose exclusivity that is absent from the source theory. The mapping retains every domain association supported by the definition-derived lexical material. In corpus application, this design requires separate reporting of unique-token coverage and domain-assignment counts because a token mapped to multiple domains can contribute more than one assignment.

3. Methods

3.1 Operational Mapping Design

This study used a descriptive, set-based design to audit an operational mapping before applying it to a literary corpus. The audit compared a definition-derived inventory with a synonym-derived sensitivity inventory while keeping the two inventories separate throughout the analysis.

3.2 Definition-Derived Lexical Template

The lexical template was extracted from Ross et al.’s (2020) five domain titles, subdomain descriptions, stated requirements, and type-of-well-being labels (pp. 474–475). Function words, auxiliaries, grammatical markers, and general relational items without substantive domain content were excluded under a predeclared content-word rule. The exclusion log retained the excluded form, its source domain, and the reason for exclusion.

For each domain, the retained lexical items were mapped to semantic tags and then deduplicated to form a domain-specific definition-derived set. When the same semantic tag

was supported by lexical material from more than one domain, it was retained in each relevant domain set rather than assigned to a single primary domain. This procedure produced five domain-specific inventories and an overall definition-derived union for subsequent set comparisons. The mapping was therefore treated as a traceable operational crosswalk from the published well-being framework to the semantic tagset, not as evidence that the two classification systems are taxonomically equivalent.

3.3 Screened Synonym-Derived Condition

The sensitivity condition was built from 51 selected definition-derived source terms. For each source term, the top-ranked synonym candidates in the Merriam-Webster online thesaurus were used as the screening pool; the thesaurus presents synonyms, similar words, and antonyms through a ranking system (Merriam-Webster, n.d.).

Up to five candidates were retained when they preserved the context-specific sense of the source term within Ross et al.'s (2020) five-domain framework (pp. 474–475). The synonym-derived inventory was kept separate from the definition-derived inventory throughout the audit because its purpose was to test the stability and marginal consequences of lexical expansion, not to redefine the source construct. Candidate synonyms were screened before the semantic annotation output was inspected, and no open-ended expansion was permitted after annotation. The frozen synonym table recorded the corresponding domain in Ross et al.'s (2020) framework, the source term, part of speech, and retained synonyms.

3.4 Processing and Set Construction

Both lexical conditions were processed in Wmatrix 7 (UCREL, 2025), a web-based tool that supports part-of-speech tagging and semantic annotation (Rayson, 2008, p. 529). The definition-derived and synonym-derived items were kept in separate input tables, each retaining the corresponding domain in Ross et al.'s (2020) five-domain framework (pp. 474–475). The exported part-of-speech tags and semantic tags were then reconciled with the input tables by domain, lexical item, and part of speech. Documented repairs to malformed spreadsheet cells were applied only from the correction log; no repair was inferred from numeric cell values. Data cleaning was conducted to remove formatting noise, documented export irregularities, and other non-substantive artifacts before duplicate domain-item-tag records were removed. The cleaned records were then used to construct the domain-specific sets and the overall union for each condition.

A lexical item appearing in more than one domain in Ross et al.'s (2020) framework was not deduplicated across domains, because its domain provenance differed. Within a given domain, however, repeated occurrences of the same full semantic tag were counted once when constructing the relevant domain-specific audit set. This deduplication rule applied only to the mapping audit; any corpus-level frequency analysis would count each tagged token as an occurrence.

3.5 Set-Based Mapping Audits

The study used a descriptive, set-based design to audit an operational mapping before applying it to a literary corpus. Two inventories were kept separate. The primary inventory contained semantic tags derived directly from Ross et al.'s

(2020) domain titles, subdomains, requirements, and well-being-type labels. The sensitivity inventory contained semantic tags generated from screened thesaurus synonyms of selected definition-derived terms. The synonym-derived inventory was used only for validation and was not merged into the primary mapping.

The lexical extraction unit was a domain-bearing content word or multiword expression drawn from the domain definitions or their screened synonyms. Let D denote the overall definition-derived tag union and S denote the overall synonym-derived tag union. For RQ1, containment was calculated as $|S \cap D| / |S|$, novelty as $|S \setminus D| / |S|$, and Jaccard similarity as $J(D, S) = |D \cap S| / |D \cup S|$. Containment estimated the proportion of synonym-derived semantic tags already represented in the definition-derived inventory, whereas novelty estimated the proportion introduced only through synonym expansion.

For RQ2, each domain-specific synonym-derived set was compared with each of the five domain-specific definition-derived sets. Alignment was assessed by whether the comparison with the corresponding domain in Ross et al.'s (2020) framework produced the highest Jaccard similarity relative to the four non-corresponding domains. For RQ3, overlap was calculated across the ten unordered pairwise comparisons among the five definition-derived domain sets. Repeated occurrences of a source word, synonym, or semantic tag within the same domain did not increase set similarity. No novel, character, or token from the literary corpus entered these set calculations.

Because the analyzed inventories were fixed, deliberately constructed sets rather than probability samples, all RQ1–RQ3 statistics were interpreted descriptively. Jaccard coefficients were used to summarize set similarity, not to test whether an observed value was statistically unusual. This descriptive application is compatible with Real and Vargas's (1996) distinction between calculating Jaccard similarity and assigning a probabilistic interpretation to whether a value is unusually high or low (pp. 380–381).

4. Results

4.1 Overall Containment and Novelty

The definition-derived mapping comprised 253 domain-tag assignments and 128 unique semantic tags. The domain-specific sets contained 54 tags in D1, 45 in D2, 68 in D3, 44 in D4, and 42 in D5. After discourse-marker and unmatched outputs were excluded, the synonym-derived condition comprised 81 domain-tag assignments and 63 unique tags: 32 in D1, 24 in D2, 9 in D3, 8 in D4, and 8 in D5.

Of the 63 unique synonym-derived tags, 44 occurred in the definition-derived union and 19 did not. Containment was therefore .698, novelty .302, and Jaccard similarity .299 (Table 1). Nearly 70% of the synonym-derived inventory was already represented in the definition-derived mapping, but synonym expansion still introduced a nontrivial 30.2% without direct support from the source definitions. The lower Jaccard coefficient is expected because the definition-derived inventory was approximately twice as large as the synonym-derived inventory. Containment measures coverage of the synonym inventory, whereas

Jaccard similarity evaluates symmetric overlap across both complete sets.

Metric	Value
Definition-derived unique tags	128
Synonym-derived unique tags	63
Shared tags	44
Novel synonym-derived tags	19
Containment	.698
Novelty	.302
Jaccard similarity	.299

Table 1: Containment and Novelty

4.2 Intended versus Cross-Domain Alignment

Same-domain Jaccard similarity was the largest value for D1, D4, and D5 (Figure 1). D1 aligned with D1 at .211, compared with a strongest cross-domain value of .163 for D3. D4 aligned with D4 at .106, compared with .060 for D2, and D5 aligned with D5 at .087, compared with .061 for D4. D2 and D3 showed the reverse ordering. D2’s same-domain similarity was .190, whereas its similarity with D4 was .214. D3’s same-domain value was .069, whereas its similarity with D1 was .086. The reversals were small in absolute terms, but they show that synonym expansion did not preserve intended-domain specificity uniformly.

The underlying intersections qualify the two reversals. D2 shared 11 tags with its own definition-derived set and 12 with D4. For D3, both the intended D3 comparison and the strongest cross-domain D1 comparison shared five tags. D1 ranked higher only because its definition-derived set was smaller than D3’s, which reduced the union denominator. The D3 result therefore does not indicate a larger shared semantic core with D1; it demonstrates the sensitivity of Jaccard rankings to set size when intersection counts are equal.

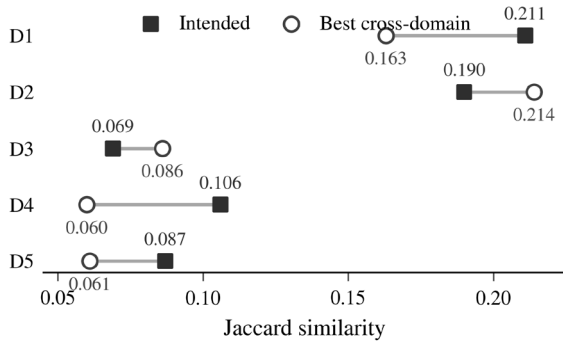


Figure 1: Intended versus Cross-Domain Alignment

4.3 Overlap Among Definition-Derived Domains

All 10 pairs of definition-derived domains overlapped, with Jaccard coefficients ranging from .210 to .328 (Figure 2). The strongest normalized overlap occurred between D2 and D4 (.328; 22 shared of 67 union tags), followed by D2-D3 (.314; 27 of 86) and D1-D3 (.312; 29 of 93). The lowest value was D1-D4 (.210; 17 of 81). No pair approached identity, but none was semantically isolated. The D2-D4 cross-domain result in the synonym audit is therefore consistent with the strongest overlap already present in the definition-derived mapping, while D3’s affinity with D1 is compatible with the third-highest definition-derived pairing.

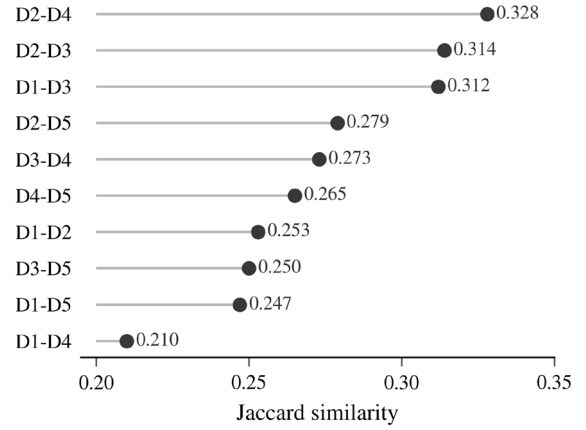


Figure 2: Definition-Derived Domain Overlap

5. Discussion

5.1 The Definition-Derived Mapping as the Primary Specification

The three audits converge on a qualified methodological decision. The definition-derived mapping covers a substantial, but not exhaustive, share of the semantic tags elicited by screened synonym expansion. Containment of .698 indicates that the retained synonyms were not arbitrary. At the same time, novelty of .302 means that merging the two inventories would add 19 tags without direct derivation from Ross et al.’s (2020) published framework. The alignment audit adds a second constraint: only three of the five synonym-derived sets were closest to their intended definition-derived domains. Taken together, these results support retaining the definition-derived mapping as the primary specification and treating the synonym-derived condition as sensitivity evidence rather than an automatic extension of the construct.

This conclusion does not require the five domains to be independent. Ross et al. (2020) describe them as interconnected, and the definition-only audit confirms overlap across every pair. A many-to-many mapping is therefore more faithful to the source framework than forced assignment of each semantic tag to a single domain. Interconnectedness, however, cannot justify unrestricted lexical expansion. Construct control is strongest when every primary tag remains traceable to the published definition and additional candidates are reported separately.

The difference between containment and Jaccard similarity is central to this interpretation. The value of .299 is not evidence that the mapping failed; it reflects the broader scope of the definition-derived inventory as well as the 19 synonym-derived tags not supported by the source definitions. The defensible reading is asymmetric: the definition-derived mapping covers much of the synonym-derived inventory, while the synonym-derived inventory samples only part of the definition-derived mapping and introduces a meaningful margin of novelty. This specification-first interpretation follows the principle that preprocessing decisions should be evaluated against the research question rather than selected after observing the most favorable output (Denny & Spirling, 2018, p. 170). It also accords with calls for transparent reporting of corpus

construction and analytical decisions (Hashimoto & Nelson, 2024, p. 60).

5.2 Domain Specificity and Cross-Domain Structure

The RQ2 findings provide partial, not universal, evidence of domain specificity. D2's proximity to D4 is substantively plausible because contribution, participation, skills, competence, and opportunity can occupy both social and educational semantic space. The D3 result requires greater restraint: equal intersections produced different Jaccard values because the candidate definition-derived sets differed in size. Neither reversal warrants relabeling a domain; both identify where domain boundaries are least distinct within the mapping.

The RQ3 results also explain why complete separation should not be expected. Every domain pair shared semantic tags, and the strongest overlaps corresponded to theoretically adjacent dimensions of social participation, learning, safety, and support. Domain-level corpus counts therefore cannot be interpreted as independent outcomes. A textual instance may legitimately contribute to more than one domain, and the reporting scheme must preserve that multiplicity without allowing repeated assignments to inflate claims about the amount of relevant language.

6. Conclusion

A theory-anchored semantic mapping can broaden corpus retrieval while preserving traceability. Because synonym expansion can introduce tags not directly supported by the source framework, the expanded inventory must be evaluated separately from the definition-derived mapping. The 128-tag definition-derived union contained 44 of the 63 synonym-derived tags, corresponding to .698 containment, .302 novelty, and .299 Jaccard similarity. Intended-domain alignment was strongest in three of five cases, while overlap among the definition-derived domains was moderate rather than negligible, ranging from .210 to .328. Together, these findings justify retaining the definition-derived inventory as the primary mapping and using synonym-derived tags as targeted sensitivity evidence.

This study provides an internally audited corpus-linguistic procedure for constructing a transparent semantic mapping from a multidimensional theoretical framework. Its purpose is to make the provenance and boundaries of semantic indicators explicit while preserving the distinction between textual indicators and the construct they operationalize.

7. Limitations

The analysis treats exact semantic tags as categorical and uses a deliberately limited synonym condition. Jaccard similarity does not represent hierarchical proximity among related tags, and unequal domain-set sizes can affect rankings. The mapping should therefore be understood as a transparent operationalization of Ross et al.'s (2020) framework rather than a complete ontology of adolescent well-being.

8. Bibliographical References

- Archer, D., Wilson, A., & Rayson, P. (2002). *Introduction to the USAS category system*. UCREL, Lancaster University. https://ucrel.lancs.ac.uk/usas/usas_guide.pdf
- Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, 26(2), 168-189. <https://doi.org/10.1017/pan.2017.44>
- Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3), 267-297. <https://doi.org/10.1093/pan/mps028>
- Hashimoto, B., & Nelson, K. (2024). Recent trends in corpus design and reporting: A methodological synthesis. *Research in Corpus Linguistics*, 12(1), 59-88. <https://doi.org/10.32714/ricl.12.01.03>
- Langford, B. H., Krauss, S. M., & Legters, L. (2021). *Investing in the well-being and well-becoming of America's young people: Recommendations for philanthropy, policy, and practice*. Youth Transition Funders Group. <https://www.ytfg.org/ytfg-publications/investing-in-the-well-being-and-well-becoming-of-americas-young-people>
- Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), 35-65. <https://doi.org/10.1111/j.1540-6261.2010.01625.x>
- McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
- Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4), 519-549. <https://doi.org/10.1075/ijcl.13.4.06ray>
- Rayson, P. (2021). *Extending the key semantic domains method beyond English corpora: Wmatrix version 5* [Conference poster]. Corpus Linguistics 2021, Limerick, Ireland.
- Rayson, P., Archer, D., Piao, S., & McEnery, A. M. (2004). The UCREL semantic analysis system. In *Proceedings of the beyond named entity recognition semantic labelling for NLP tasks workshop, Lisbon, Portugal, 2004* (pp. 7-12).
- Real, R., & Vargas, J. M. (1996). The probabilistic basis of Jaccard's index of similarity. *Systematic Biology*, 45(3), 380-385. <https://doi.org/10.1093/sysbio/45.3.380>
- Ross, D. A., Hinton, R., Melles-Brewer, M., Engel, D., Zeck, W., Fagan, L., Herat, J., Phaladi, G., Imbago-Jacome, D., Anyona, P., Sanchez, A., Damji, N., Terki, F., Baltag, V., Patton, G., Silverman, A., Fogstad, H., Banerjee, A., & Mohan, A. (2020). Adolescent well-being: A definition and conceptual framework. *Journal of Adolescent Health*, 67(4), 472-476. <https://doi.org/10.1016/j.jadohealth.2020.06.042>
- Widdowson, H. G. (2008). The novel features of text: Corpus analysis and stylistics. In A. Gerbig & O. Mason (Eds.), *Language, people, numbers: Corpus linguistics and society* (pp. 293-304). Rodopi.

9. Language Resource References

- Archer, D., Wilson, A., & Rayson, P. (2002). *Introduction to the USAS category system* [Semantic tagset documentation]. UCREL, Lancaster University. https://ucrel.lancs.ac.uk/usas/usas_guide.pdf
- Merriam-Webster. (n.d.). *The Merriam-Webster thesaurus* [Online thesaurus]. Retrieved June 8, 2026, from <https://www.merriam-webster.com/thesaurus>

UCREL. (2025). *Wmatrix corpus analysis and comparison tool* (Version 7) [Computer software]. Lancaster University. <https://ucrel-wmatrix7.lancaster.ac.uk/>