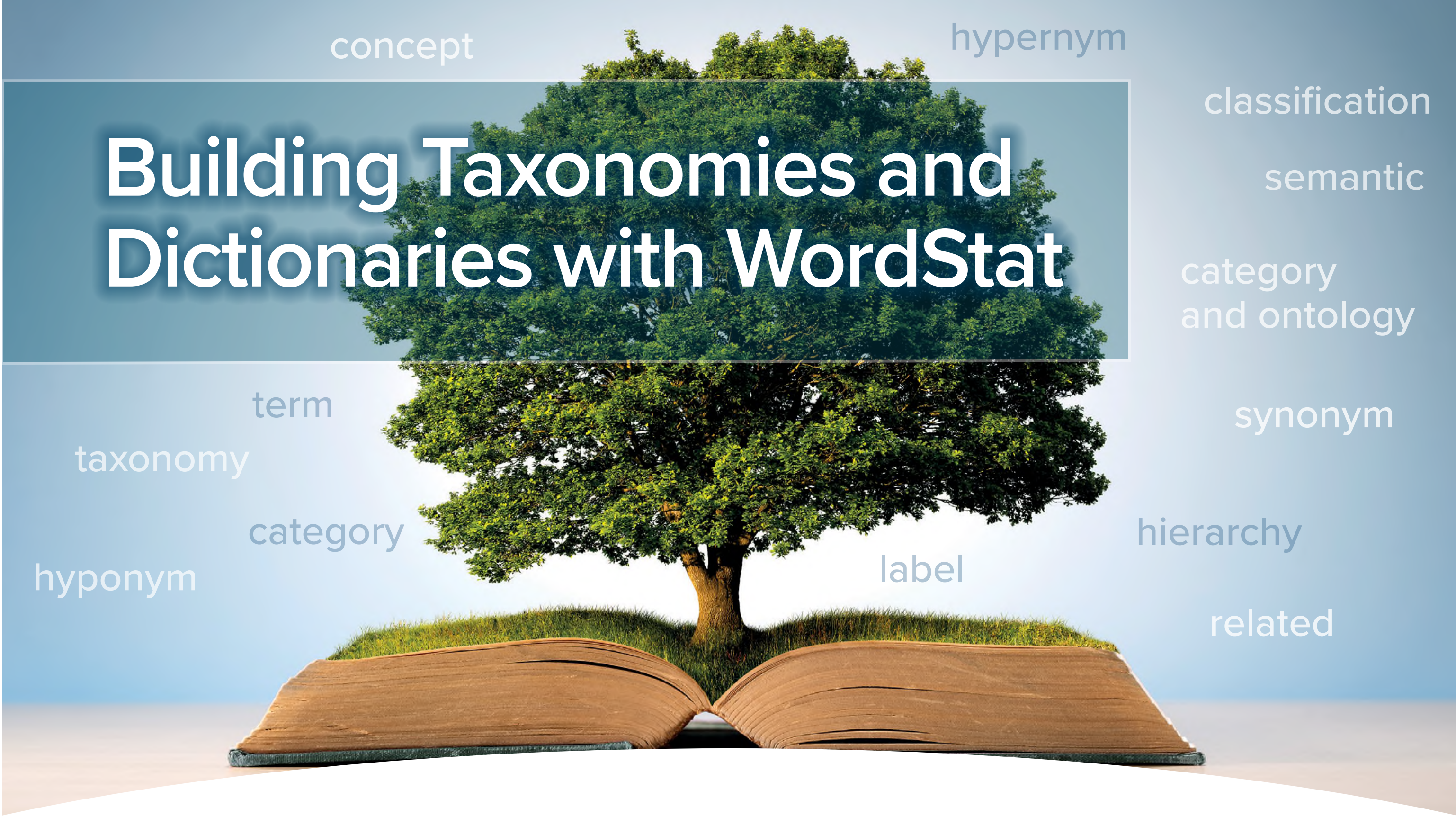
Building Taxonomies and Dictionaries with WordStat



The Value of Content Analysis Dictionaries

Taxonomies and content analysis dictionaries are structured vocabularies that organize words and phrases into meaningful categories, facilitating systematic text coding and analysis. Sometimes called lexicons, or taxonomies when organized hierarchically, these tools map words, phrases, or patterns to well-defined conceptual categories. They enable researchers to measure abstract constructs such as emotions, values, attitudes, themes, or cognitive strategies across large volumes of text. Unlike inductive techniques (topic modeling, neural networks), dictionary-based approaches follow a deductive, rule-based logic where classification criteria are explicitly defined by the analyst. In this sense, content analysis dictionaries constitute a form of **rule-based artificial intelligence**, deriving meaning through carefully designed linguistic and contextual rules.

This rule-based method offers several benefits: it ensures **transparency,** since each classification decision is fully visible and auditable; it guarantees **replicability**, enabling consistent application across datasets or over time; and it promotes **conceptual clarity**, by aligning categories with theoretical frameworks or operational definitions. These qualities make it indispensable in domains requiring precision and interpretability, such as social science, market research, education, public policy, and risk analysis.

Dictionaries and taxonomies are used in numerous fields. In **psychology**, they quantify cognitive styles, emotional expression, or personality traits. In **political science**, they help track ideology, framing strategies, or rhetorical shifts in party platforms and speeches. **Businesses** rely on taxonomies to monitor customer satisfaction, brand sentiment, or feedback themes. **Risk assessment and security experts** may use dictionaries to analyze incident and maintenance reports to detect threats, identify recurring issues, or support predictive maintenance strategies. In **education**, they help detect references to skills, competencies, or learning outcomes in student essays or curriculum guidelines. In each of these contexts, dictionaries make it possible to scale up qualitative insight, producing quantitative indicators while preserving interpretability.

PROVALIS
RESEARCH

Dictionary Construction in WordStat

Yet building these dictionaries is not trivial. Analysts must contend with challenges such as the **linguistic variability**, the many ways a single idea can be expressed, as well as **word polysemy**, where the same word can convey multiple meanings. **Misspellings**, lexical variation, and **evolving terminology** further complicate the task. WordStat was designed to support users throughout this process, providing a comprehensive suite of tools to develop, validate, and refine content analysis dictionaries using both linguistic resources and advanced analytics. Its approach to content dictionary development recognizes that dictionary building is both a linguistic and an analytical task. The software offers tools to build dictionaries from scratch or adapt existing ones to specific contexts. Users can define simple keyword lists or construct more complex hierarchical structures with multilevel categories and embedded proximity rules. Users can also import domain-specific taxonomies stored in **SKOS format (RDF or TTL files)**, allowing them to build on existing resources. Beyond manual entry, WordStat includes an array of features to accelerate the process, such as phrase extraction, spelling correction, and numerous suggestion tools based on statistical, semantic, and contextual analyses.

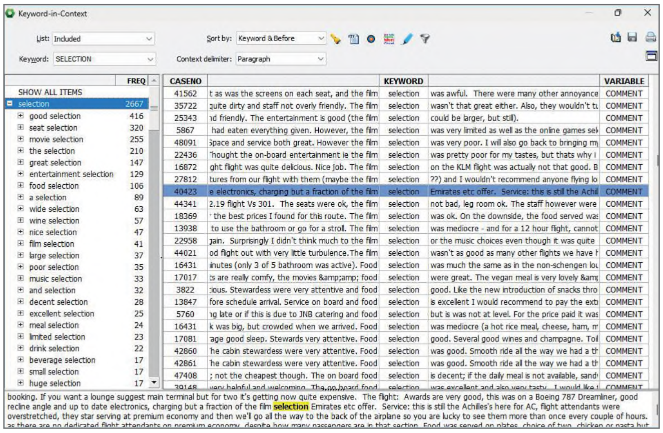
For example, WordStat excels in identifying high-frequency words and multi-word phrases, letting users explore, validate, and categorize them with ease. Analysts can explore these items interactively, search for variations, examine their contexts, and rapidly assign them to appropriate categories. To help identify variants and semantically related expressions, WordStat provides several complementary techniques, including the integration of several lexical resources (**WordNet**, two English **thesauri** and, for multilanguage coverage, thesauri in seven additional languages). It also uses **corpus-trained word embeddings**, which allow it to suggest semantically related words and phrases in virtually any language (as long as sufficient text is available). Unlike generic synonym lists, WordStat's suggestions are context-sensitive and corpus-aware, displaying only related terms that actually occur in the corpus, along with their frequencies. This targeted approach allows one to efficiently achieve adequate coverage and avoids cluttering dictionaries with synonyms that are unlikely to be used in the data. Users may also choose to view a more comprehensive list of related terms to ensure the generalizability of the dictionary to other datasets or applications.

Multi-word expressions are managed as meaningful linguistic units with flexible pattern definitions. Users can perform searches and **wildcard matching** to find and group related variants or rely on embedding-based similarity metrics to detect near-equivalents. This is particularly useful when dealing with idiomatic expressions, domain-specific terminology, or evolving language.

Spelling errors can also represent a significant source of **false negatives** in dictionary-based content analysis, particularly when analyzing unstructured or user-generated content. To mitigate this, WordStat includes a unique, **high-speed spell-checking** engine capable of scanning millions of words in just a few seconds. It automatically identifies potential spelling mistakes in the corpus and automatically applies high-confidence corrections. It will even **automatically recognize unknown terms**, such as proper nouns, neologisms, or technical terms, and identify their misspelled forms, a capability not found in conventional spell-checkers. The correction list can be fully reviewed and edited, allowing users to inspect or adjust suggested corrections and manually add additional substitutions when needed. A separate feature identifies additional low-confidence spelling issues for manual review. Spelling dictionaries are available in **37 languages**, and include dedicated legal and medical dictionaries.

Achieving Precision Through Disambiguation

To refine and validate dictionaries and resolve ambiguity, WordStat provides an advanced keyword-in-context or **KWIC concordancer**.



List: Included

Sort by: Keyword & Before

Keyword: SELECTION

Context delimiter: Paragraph

	FREQ
SHOW ALL ITEMS	
selection	2667
good selection	416
seat selection	320
movie selection	255
the selection	210
great selection	147
entertainment selection	129
food selection	106
a selection	89
wide selection	63
wine selection	57
nice selection	47
film selection	41
large selection	37
poor selection	35
music selection	33
and selection	32
decent selection	28
excellent selection	25
meal selection	24
limited selection	23
drink selection	22
beverage selection	17
small selection	17
huge selection	17

CASENO		KEYWORD		VARIABLE
41562	t as was the screens on each seat, and the film	selection	was awful. There were many other annoyance	COMMENT
35722	quite dirty and staff not overly friendly. The film	selection	wasn't that great either. Also, they wouldn't tu	COMMENT
25343	ıd friendly. The entertainment is good (the film	selection	could be larger, but still).	COMMENT
5867	had eaten everything given. However, the film	selection	was very limited as well as the online games sel	COMMENT
48091	Space and service both great. However the film	selection	was very poor. I will also go back to bringing my	COMMENT
22436	Thought the on-board entertainment ie the film	selection	was pretty poor for my tastes, but thats why i	COMMENT
16872	ght flight was quite delicious. Nice job. The film	selection	on the KLM flight was actually not that good. B	COMMENT
27812	tures from our flight with them (maybe the film	selection	??) and I wouldn't recommend anyone flying lo	COMMENT
40423	e electronics, charging but a fraction of the film	selection	Emirates etc offer. Service: this is still the Achil	COMMENT
44341	2.19 flight Vs 301. The seats were ok, the film	selection	not bad, leg room ok. The staff however were	COMMENT
18369	the best prices I found for this route. The film	selection	was ok. On the downside, the food served was	COMMENT
13938	to use the bathroom or go for a stroll. The film	selection	was mediocre - and for a 12 hour flight, cannot	COMMENT
22958	gain. Surprisingly I didn't think much to the film	selection	or the music choices even though it was quite	COMMENT
44021	od flight out with very little turbulence.The film	selection	wasn't as good as many other flights we have h	COMMENT
16431	inutes (only 3 of 5 bathroom was active). Food	selection	was much the same as in the non-schengen lou	COMMENT
17017	ts are really comfy, the movies &amp; food	selection	were great. The vegan meal is very lovely &am	COMMENT
3822	ious. Stewardess were very attentive and food	selection	good. Like the new introduction of snacks thro	COMMENT
13847	fore schedule arrival. Service on board and food	selection	is excellent I would recommend to pay the extr	COMMENT
5760	ng late or if this is due to JNB catering and food	selection	but is was not at level. For the price paid it was	COMMENT
16431	k was big, but crowded when we arrived. Food	selection	was mediocre (a hot rice meal, cheese, ham, m	COMMENT
17081	age good sleep. Stewards very attentive. Food	selection	good. Several good wines and champagne. Toil	COMMENT
42860	he cabin stewardess were very attentive. Food	selection	was good. Smooth ride all the way we had a th	COMMENT
42861	he cabin stewardess were very attentive. Food	selection	was good. Smooth ride all the way we had a th	COMMENT
47408	; not the cheapest though. The on board food	selection	is decent; if the daily meal is not available, sand	COMMENT
39148	very helpful and welcoming. The on board food	selection	was excellent and also very tasty. I would like t	COMMENT

booking. If you want a lounge suggest main terminal but for two it's getting now quite expensive. The flight: Awards are very good, this was on a Boeing 787 Dreamliner, good recline angle and up to date electronics, charging but a fraction of the film selection Emirates etc offer. Service: this is still the Achilles's here for AC, flight attendants were overstretched, they star serving at premium economy and then we'll go all the way to the back of the airplane so you are lucky to see them more than once every couple of hours. as there are no dedicated flight attendants on premium economy, despite how many passengers are in that section. Food was served on plates, choice of two, chicken or pasta but

Figure 1. WordStat's KWIC concordancer displays keyword contexts with hierarchical grouping and frequency ranking for precise dictionary validation and disambiguation.

This KWIC feature displays occurrences of a given word or phrase within their immediate context, allowing users to assess how they are actually being used in the text. The list can be sorted based on the words that precede or follow the keyword, making it easier to identify consistent patterns or divergent uses. A panel on the left side of the interface automatically groups similar context sequences in a hierarchical tree and ranks them by frequency, helping analysts focus on the most frequent usage cases. This allows the users to quickly assess the achieved precision of the target term or identify false positives and add them to a different dictionary category or to a custom exclusion list. In case of highly ambiguous terms, it may also be used to identify and select true positive phrases. This is a crucial feature to refine dictionaries and achieve high precision in measurement.

Such high accuracy is only possible thanks to WordStat's unique built-in **hierarchical matching logic** (see figure 2), which prevents double-counting by treating overlapping items as mutually exclusive and consistently prioritizing more specific entries over less specific ones. Specifically, longer phrases take precedence over shorter ones; phrases are favored over individual words; case-sensitive matches override case-insensitive ones (e.g., March vs. march); whole words are matched before wildcard patterns; and longer patterns always have priority over shorter ones. This hierarchical processing is essential for **disambiguation**, allowing users to include ambiguous terms in their dictionaries while also specifying phrases or patterns that clarify or exclude unintended meanings. Many text analysis tools lack this precedence structure, leading to inflated or inconsistent category counts.

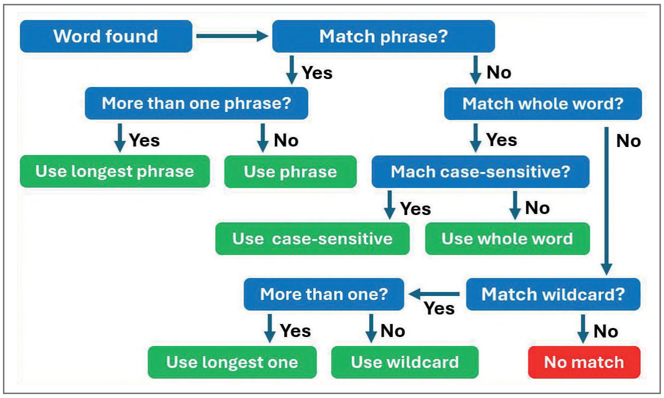


Figure 2. Hierarchical matching logic preventing double-counting and ensuring accurate disambiguation.

When phrase selection is insufficient for disambiguation, WordStat's advanced **proximity rules** can be used, enabling precise, context-aware coding. These rules combine **Boolean logic** (AND, OR, NOT) with **proximity operators** such as BEFORE, AFTER, NEAR, and their negations. They allow users to specify that terms or categories must co-occur in the same structural unit, whether a **document, paragraph,** or **sentence**, and be separated by no more than a user-defined number of words. Those proximity rules are useful to handle negations in sentiment analysis, disambiguate polysemous words through contextual cues, detect conditional or hedged statements. And by capturing the co-occurrence of key elements in context, proximity rules are crucial for detecting events or actions, offering a parsimonious alternative to enumerating the many possible phrasings that may be used to express them.

WordStat also allows dictionary items to be **weighted** individually. This supports scoring systems, where different terms may carry different analytical importance. It also can function as a form of probabilistic disambiguation, where ambiguous terms are proportionally distributed to different concepts based on usage.

Taxonomies in WordStat are saved as comprehensive categorization models (*.wmodel files), encompassing the categorization dictionary, a custom stop word list, previously identified spelling corrections, and all preprocessing settings (stemming, lemmatization, tokenization, etc.) configured by the user and necessary to ensure reliable replication across datasets. These models are **fully portable**, reusable across projects, and preserve complete configuration integrity. Moreover, a **software development kit (SDK)** allows seamless integration of these models into third-party applications, enabling on-the-fly text processing and **real-time analysis** within data collection systems, thereby supporting embedded analytics and operational deployment.

Exploratory Text Analysis Tools

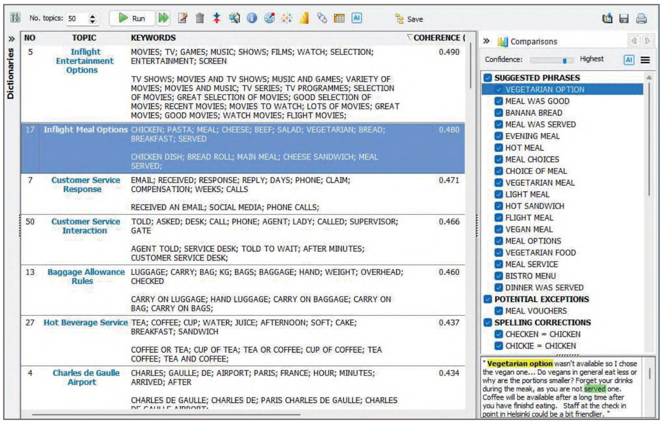
Exploratory text analysis techniques offer powerful ways to uncover underlying structures and patterns within large text corpora. By automatically grouping related terms and concepts based on their distribution and co-occurrence, these methods help analysts detect emerging themes, identify semantically coherent clusters, and reveal relationships that may not be immediately obvious through manual inspection. This data-driven insight provides empirical guidance for developing and refining taxonomies, ensuring that classification categories effectively capture the nuances present in the data.

WordStat includes several exploratory tools, including:

- **Topic modeling**
- **Hierarchical clustering**
- **Multidimensional scaling**
- **Named entity recognition**
- **Correspondence analysis**
- **Crosstabulation**
- **Deviation tables**

These techniques allow users to uncover latent themes, visualize associations among terms and categories, and identify differences across sources or over time. They support the discovery of relevant concepts, help define taxonomy structures, and highlight gaps in existing dictionaries, bridging open-ended exploration and structured content analysis.

One of WordStat's standout features is its **topic modeling** module (see figure 3), which produces highly coherent, stable, and interpretable topics that outperform widely used methods like Latent Dirichlet Allocation (LDA) as well as recent word-based or sentence-based embedding-based approaches such as BERTopic. The model **automatically incorporates relevant multi-word phrases** alongside single words, enhancing both coherence and interpretability. It also suggests related phrases to expand topic coverage and flags those containing potential false positives, facilitating disambiguation and refinement. Topics are named automatically using statistical criteria, with **generative AI** optionally improving these names and organizing topics hierarchically to create clearer thematic structures. These complete topic structures, including hierarchical organization, names, and phrases, can be converted into a content analysis dictionary with a single button click.



NO	TOPIC	KEYWORDS	COHERENCE (
5	Inflight Entertainment Options	MOVIES; TV; GAMES; MUSIC; SHOWS; FILMS; WATCH; SELECTION; ENTERTAINMENT; SCREEN TV SHOWS; MOVIES AND TV SHOWS; MUSIC AND GAMES; VARIETY OF MOVIES; MOVIES AND MUSIC; TV SERIES; TV PROGRAMMES; SELECTION OF MOVIES; GREAT SELECTION OF MOVIES; GOOD SELECTION OF MOVIES; RECENT MOVIES; MOVIES TO WATCH; LOTS OF MOVIES; GREAT MOVIES; GOOD MOVIES; WATCH MOVIES; FLIGHT MOVIES;	0.490
17	Inflight Meal Options	CHICKEN; PASTA; MEAL; CHEESE; BEEF; SALAD; VEGETARIAN; BREAD; BREAKFAST; SERVED CHICKEN DISH; BREAD ROLL; MAIN MEAL; CHEESE SANDWICH; MEAL SERVED;	0.480
7	Customer Service Response	EMAIL; RECEIVED; RESPONSE; REPLY; DAYS; PHONE; CLAIM; COMPENSATION; WEEKS; CALLS RECEIVED AN EMAIL; SOCIAL MEDIA; PHONE CALLS;	0.471
50	Customer Service Interaction	TOLD; ASKED; DESK; CALL; PHONE; AGENT; LADY; CALLED; SUPERVISOR; GATE AGENT TOLD; SERVICE DESK; TOLD TO WAIT; AFTER MINUTES; CUSTOMER SERVICE DESK;	0.466
13	Baggage Allowance Rules	LUGGAGE; CARRY; BAG; KG; BAGS; BAGGAGE; HAND; WEIGHT; OVERHEAD; CHECKED CARRY ON LUGGAGE; HAND LUGGAGE; CARRY ON BAGGAGE; CARRY ON BAG; CARRY ON BAGS;	0.460
27	Hot Beverage Service	TEA; COFFEE; CUP; WATER; JUICE; AFTERNOON; SOFT; CAKE; BREAKFAST; SANDWICH COFFEE OR TEA; CUP OF TEA; TEA OR COFFEE; CUP OF COFFEE; TEA COFFEE; TEA AND COFFEE;	0.437
4	Charles de Gaulle Airport	CHARLES; GAULLE; DE; AIRPORT; PARIS; FRANCE; HOUR; MINUTES; ARRIVED; AFTER CHARLES DE GAULLE; CHARLES DE; PARIS CHARLES DE GAULLE; CHARLES DE GAULLE AIRPORT;	0.434

☑ SUGGESTED PHRASES
- ✓ VEGETARIAN OPTION
- ☑ MEAL WAS GOOD
- ☑ BANANA BREAD
- ☑ MEAL WAS SERVED
- ☑ EVENING MEAL
- ☑ HOT MEAL
- ☑ MEAL CHOICES
- ☑ CHOICE OF MEAL
- ☑ VEGETARIAN MEAL
- ☑ LIGHT MEAL
- ☑ HOT SANDWICH
- ☑ FLIGHT MEAL
- ☑ VEGAN MEAL
- ☑ MEAL OPTIONS
- ☑ VEGETARIAN FOOD
- ☑ MEAL SERVICE
- ☑ BISTRO MENU
- ☑ DINNER WAS SERVED

☑ POTENTIAL EXCEPTIONS
- ☑ MEAL VOUCHERS

☑ SPELLING CORRECTIONS
- ☑ CHECKEN = CHICKEN
- ☑ CHICKIE = CHICKEN

" **Vegetarian option** wasn't available so I chose the vegan one... Do vegans in general eat less or why are the portions smaller? Forget your drinks during the meak, as you are not served one. Coffee will be available after a long time after you have finishd eating.   Staff at the check in point in Helsinki could be a bit friendlier. "

Figure 3. WordStat's topic modeling interface showing coherent topics with automatic phrase suggestions, false positive detection, and integrated spell-checking capabilities.

Complement to Ontology Tools

For researchers working with formal ontology development and graph databases, WordStat provides essential empirical validation capabilities that complement semantic modeling tools. While ontology editors excel at formal knowledge representation using standards like OWL and SKOS, powering inference engines and ensuring interoperability across structured datasets, they typically lack the empirical text analysis tools needed to validate concepts against real-world language usage.

WordStat bridges this gap through corpus-based frequency analysis, KWIC contextual validation, and sophisticated disambiguation logic. The workflow is bidirectional: WordStat can serve as an empirical front-end for ontology development, supporting exploratory discovery, hypothesis generation, and iterative refinement before formalizing taxonomies. Conversely, existing ontological structures can be imported as starting dictionaries for corpus-based validation and refinement.

While formal taxonomies provide essential conceptual frameworks, they can be too abstract for precise recognition in complex real-world texts. WordStat enhances their practical applicability by grounding categories in empirical language usage and employing context-aware disambiguation. Results directly inform RDF schema design, SKOS development, and relationship modeling in graph databases.

Conclusion

By combining transparent, rule-based methods with advanced exploratory analytics, WordStat complements modern AI and NLP techniques, offering users a balanced approach to both theory-driven and data-driven text analysis. It offers a rich, transparent, and adaptable environment for developing content analysis dictionaries and taxonomies. Whether the goal is theory-driven measurement, cross-linguistic analysis, risk monitoring, or conceptual discovery, WordStat delivers the tools to move from unstructured text to structured insight, efficiently, rigorously, and with confidence.

PROVALIS
RESEARCH

To schedule a web demo or for more information on our products,
contact us: TOLL FREE 1 855 355-5252 or 1 514 899-1672
sales@provalisresearch.com, provalisresearch.com