# Word Midas Powered by StringNet: Discovering Lexicogrammatical Constructions *in Situ*[1]

**David Wible**
National Central University
No.300, Jhongda Rd. Jhongli City,
Taoyuan County 32001, Taiwan
wible@stringnet.org

**Nai-Lung Tsao**
Tamkang University
No.151, Yingzhuan Rd., Tamsui Dist.,
New Taipei City 25137, Taiwan
beaktsao@stringnet.org[2]

## Abstract

Adult second language learners face the daunting but underappreciated task of mastering patterns of language use that are neither products of fully productive grammar rules nor frozen items to be memorized. Word Midas[3], a web browser extension, targets this uncharted territory of lexicogrammar by detecting multiword tokens of lexicogrammatical patterning in real time *in situ* within the noisy digital texts from the user's unscripted web browsing or other digital venues. The language model powering Word Midas is StringNet, a densely cross-indexed navigable network of one billion lexicogrammatical patterns of English. These resources are described and their functionality is illustrated with a detailed scenario.

## 1 Background

Some of the most persistent yet underappreciated challenges in learning a second language are to be found not within the purview of highly productive grammar rules nor in particular lexical items whose behavior is well described by those rules and by traditional dictionary entries. They lie rather in a vast and poorly charted middle ground between items and rules, a territory of semi-productive and lexically picky patterns, what Halliday referred to as "lexis as most delicate grammar" (1961) and others call lexico-grammatical constructions (Kay, 1997). In addition to escaping coverage in traditional knowledge resources, this cline of lexicogrammar often skirts the awareness of language teachers. In the absence of reference books and teachers, children learn these elusive and variegated patterns in their first language by some combination of immersive encounters with tokens of language in use and their uncanny acumen in distilling abstract patterns from these tokens. Rare is the adult second language learner, however, who shares the child's urgency and immersion in target language input or the child's capacity for detecting patterns from it. The tools we present in this paper address basic challenges posed to adult learners and their teachers by this intermediate territory of lexicogrammar.

As an example of the challenge, the string *a wide range of issues* conforms to maximally general rules of English grammar, and so that would seem adequate to a command of this expression. This is misleading, however. Simpson-Vlack and Ellis (2010) identified the 4-gram *a wide range of* as the top-ranked item on their corpus-derived academic formula list. This shows that the co-occurrence of the words in that 4-gram is the result of more than simply general combinatorial rules of syntax and that it deserves more attention from learners than grammar alone would indicate. Our point is that here, as in so many patterns that pervade language, the expression is neither a frozen, one-off item nor simply a product of maximally general rules combining words. Rather, *a wide range of* is part of a tight nexus of limited but inter-related variations: *a [wide/broad/whole/vast] range of*; *a wide [range/variety/array/spectrum] of*.



[1] This research was supported by Taiwan's Ministry of Science and Technology, grant #105-2511-S-008-008.

[2] The work reported here was first done when Nai-Lung Tsao worked at National Central University.

[3] Word Midas can be downloaded for free from Google Chrome Store.



*Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 21–24, Osaka, Japan, December 11-17 2016.

The problem we address is this: With current resources, it is virtually impossible for situated learners to get any sense of which strings in their input are instantiations of lexicogrammatical patterning and, for those that are, what variations on those strings are possible. The system we present, called Word Midas, detects tokens of such patterns in noisy digital text of the readers' choice and shows the paradigmatic variation that is not present in the text but which constitutes the possibilities available there that are part of a mature language users grasp of the language. In what follows we describe the components of the system and illustrate its workings with a detailed scenario.

## 2 The System

The system has two basic components: (1) an existing, corpus-derived English language model called StringNet (nav4.stringnet.org), consisting of over one billion unique lexicogrammatical patterns; (2) a web browser plugin tool called Word Midas, which detects multiword strings in digital text that are instantiations of any of the lexicogrammatical patterns from the language model (StringNet).

### 2.1 The Language Model: StringNet

StringNet (Wible and Tsao, 2010; 2011) contains not only n-grams extracted from the British National Corpus (BNC) such as the trigram *as good as,* but also what Wible and Tsao (2010) call "hybrid n-grams," that is, more abstract n-grams where grams can include part-of-speech categories, for example, *as [adj] as*. The gram types that can compose hybrid n-grams fall into one of four levels of abstraction: word form (*made*, *makes*), lemma (**make**, subsuming *make*, *makes*, *made*, *making*), detailed part of speech (V-past), and rough part of speech (V). StringNet is a relational network in which the one billion unique hybrid n-grams are those from 2 to 6 grams in length instantiated in BNC with 5 or more tokens. These are cross-indexed for subsumption and inclusion relations. Thus, the n-gram *as good as* is indexed to its more abstract counterpart *as [adj] as* and to its longer counterparts, for example *as good as his word*; *as good as [poss pro] word*; and ***be*** *as good as his word*. Slots in hybrid n-grams differ in how open or selective they are in the words that can appear in them. The selective slots can be detected computationally within the more than one billion hybrid n-grams in StringNet and have served as contextual clues flagging slots of semantically similar words, what Tsao and Wible (2013) call "constructional selection."

### 2.2 The *In Situ* Tool: Word Midas

The basic conception of a browser-based tool that identifies multiword patterns in real time within texts that the user freely browses was first implemented in a tool called Collocator that detected two-word collocations (Wible et al., 2011). In the present paper, Word Midas extends the conception to the immensely more complex challenge of detecting lexicogrammatical patterns. Recall a fundamental challenge which language input poses to learners and which Word Midas aims to address: the tokens of word strings that a user encounters in input do not directly signal the lexico-grammatical pattern(s) that they betoken. This severely limits the value of such tokens and such encounters for learning unfamiliar patterns of the language behind them. Word Midas provides a reader in real time with links from the tokens in a text to the patterns they instantiate.

We illustrate the workings of Word Midas through a scenario for its use. This involves a person browsing web content with Word Midas installed on their browser as a Chrome extension or a plug-in. The user can select any word or string of words found in that web page in order to discover whether it is used there as part of a pattern and if so what variations are possible that are not there in that text. Figure 1 shows a webpage where the user has paused at the string "In a region where many people eat chocolate on a daily basis…" and wonders about the string *on a daily basis* here or simply about one of the words in that string as used here. Word Midas is activated from a context menu that appears by right-clicking on any word in that string. Activating Word Midas from that context menu launches a search for patterns that include the selected word and fall within a nine-word window of context surrounding it. The results are listed in a popup as shown in Figure 1.

Word Midas results are derived in three steps. First, string matching identifies all n-grams that include the user-selected word and match the context of that word in the text where the user found it. Then, an adapted edit-distance algorithm identifies abstract hybrid n-grams that also describe n-grams identified in the first step (see Wible and Tsao 2009 on this adaptation of edit distance). Third, pruning and ranking

algorithms evaluate these patterns using weighted features (including edit distance scores) that assess two competing requirements: their closeness of match to the target text and their productivity in wider use (see Wible and Tsao 2010 for details on this pruning and ranking of StringNet patterns). To see the significance of this, we return to the example.

The popup in Figure 1 lists the hybrid n-gram patterns that describe strings found surrounding the selected word and, for each pattern, gives its frequency in BNC and links to a concordance of sentences containing that pattern. In this case, the top three patterns listed (with their frequencies) are: *[noun] on a daily basis* (60); *[noun sg] on a daily basis (35)*; *on a daily basis* (150). The prominent frequency of the third pattern (*on a daily basis*) invites further exploration.

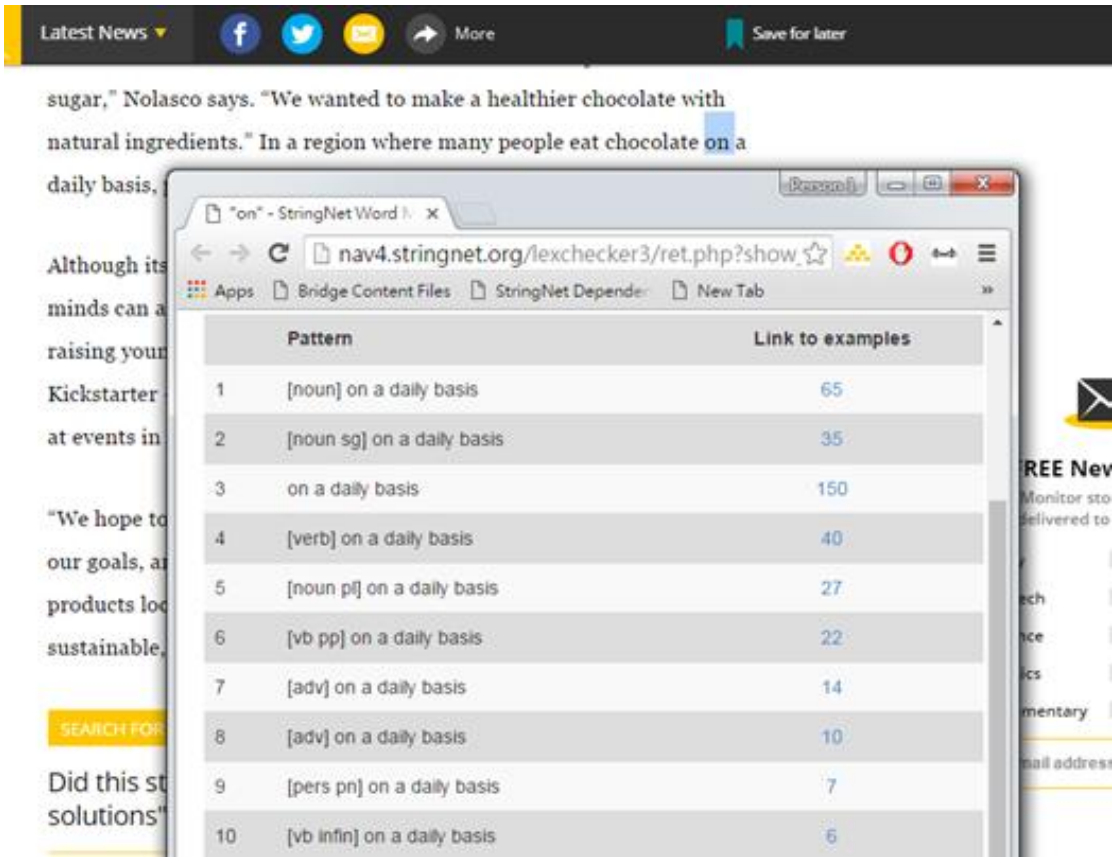

Figure 1. Patterns found by Word Midas from webpage text.

From here, Word Midas helps answer key question about lexicogrammatical patterning that an unaided reader cannot: What variations of this string are possible? Clicking on any word in the pattern *on a daily basis* where it appears in the popup list shows the user the paradigm of all words attested in that slot, indicating whether the word there is the only option or replaceable, and if replaceable, whether variation is relatively open or restricted. This can be done for each word in the sequence. In this case, the first word in *on a daily basis*, the preposition *on*, is shown to be the only attested word in its slot as is the indefinite article *a* which follows it. In turn, the noun *basis* in this frame, while replaceable, is hands down the most likely word in its position; its frequency there is 150 while the second most frequent noun, *rate*, has a frequency of only three. The limited substitutability in this four-word sequence for these three slots --*on*, *a*, and *basis*--demonstrate Sinclair's "idiom principle," that is, they are cases of tight limitations on lexical choice far more restrictive than grammar can account for. Turning to the adjective *daily* in this sequence, however, we discover a different phenomenon. Word Midas shows that the word *daily* in the sequence *on a daily basis* is occupying a slot that exhibits a dramatically wider variation than any of the other slots in this string, with 469 different adjectives attested in that position: *on a [adj] basis*. This can be seen in the paradigm for that slot shown partially in Figure 2. Thus, we can discover that *on a daily basis* is just one member in a family of variations, and that the locus of variation is concentrated in the [adj] slot.

*on a [daily/regular/voluntary/day-to-day/part-time/temporary…] basis*

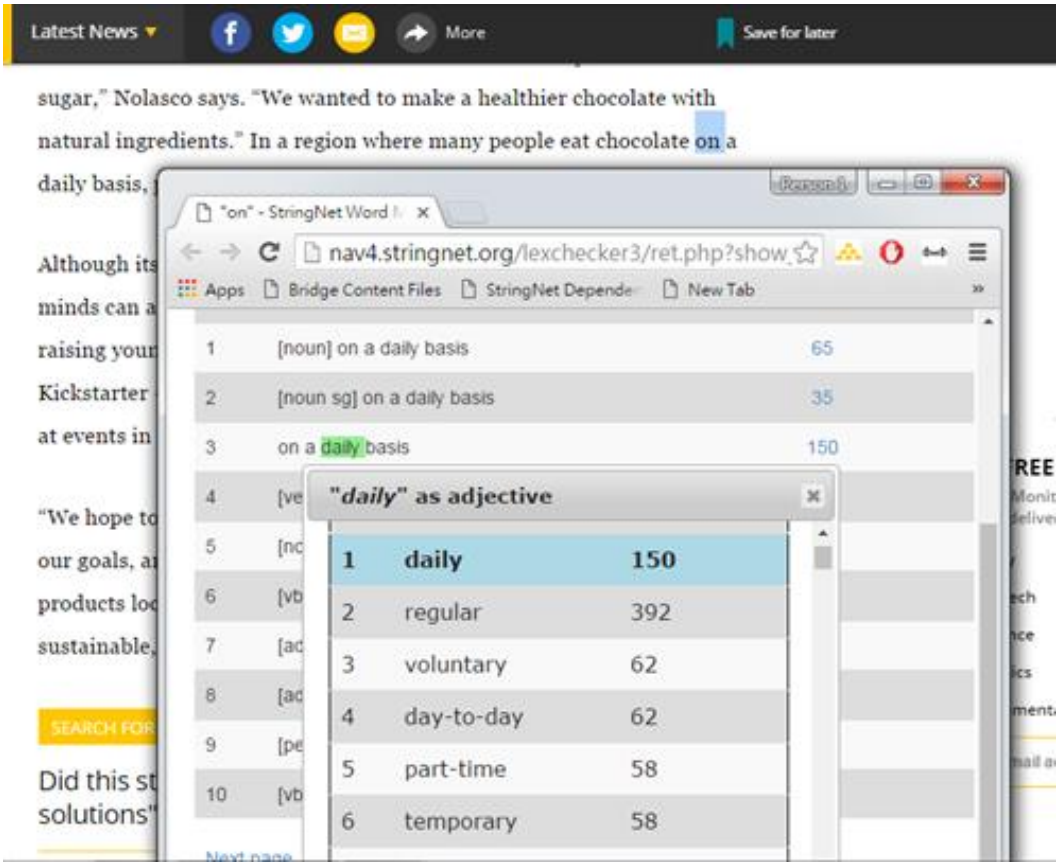


Figure 2. Paradigm for *daily* slot in *on a daily basis.*

## 3 Conclusion

There are some major points to note from this extended example. The relative freedom exhibited in the slot occupied by *daily* in the string *on a daily basis* and the corresponding lack of freedom in its co-occurring slots are symptomatic of the regularity and idiomaticity that comingle in lexicogrammatical constructions. This mix of features is undetectable from simple encounters with multiword tokens in text and unrepresented in dictionaries and grammar references. Access to these features requires access to the paradigmatic dimension of input. That access is what StringNet and Word Midas aim to provide to situated, unscripted users in real time for any part of noisy text they wish to explore.

## References

M.A.K. Halliday. 1961. Categories of the theory of grammar. *Word*, 17(3): 241-292.

Paul Kay. 1997. *Words and the Grammar of Context*. Stanford: Center for the Study of Language and Information Publications, Stanford, California.

Anne L.-E Liu, David Wible and Nai-Lung Tsao. 2011. A Browser-based Approach to Incidental Individualization of Vocabulary Learning. *Journal of Computer Assisted Learning,* 27: 540-543.

Rita Simpson-Vlach and Nick C. Ellis. 2010. An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics*, 31(4): 487-512.

Nai-Lung Tsao and David Wible. 2009. A Method for Unsupervised Lexical Error Detection and Correction. *Proceedings of the NAACL Workshop on Innovative Uses of NLP for Building Educational Applications,* 51-54.

Nai-Lung Tsao and David Wible. 2013. Word Similarity Using Constructions as Contextual Features. Joint Symposium on Semantic Processing, Trento, Italy.

David Wible and Nai-Lung Tsao. 2010. StringNet as a Computational Resource for Discovering and Investigating Linguistic Constructions. *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics*, 25-31.

David Wible and Nai-Lung Tsao. 2011. Towards a New Generation of Corpus-derived Lexical Resources for Language Learning. In F. Meunier (ed.) *A Taste of Corpora*. John Benjamins, Amsterdam.