Hype Intensity Lexicon in Biomedical Research

Dipesh Satav[1] Neil Millar[1] Bojan Batalo[2] Erica K. Shimomoto[2] Ryosuke L. Ohniwa[1]
[1]National Institute of Advanced Industrial Science and Technology (AIST) [2]University of Tsukuba
dipesh@cvlab.cs.tsukuba.ac.jp millar.neil@u.tsukuba.ac.jp
{bojan.batalo,kidoshimomoto.e}@aist.go.jp ohniwa@md.tsukuba.ac.jp

Abstract

Promotional language (“hype”) is increasingly common in biomedical research reporting. Adjectives such as *groundbreaking* can engage readers, but risk undermining objectivity. Detecting such language requires distinguishing degrees of promotional intensity (e.g., *new* < *novel* < *revolutionary*), yet no such graded resource exists. We present **Hyplex**, an intensity-scaled lexicon of 303 promotional adjectives attested in biomedical writing across eight evaluative domains (e.g., IMPORTANCE, NOVELTY, RIGOUR). Ratings were obtained through Best–Worst Scaling (BWS) with human participants evaluating adjectives for promotional strength within scientific research reporting, showing high internal consistency.

1 Introduction

In scientific writing, authors may select to use language that promotes a favourable evaluation of their work, e.g., importance may be described in absolute terms (*imperative*), novelty sensationalised (*revolutionary*), scale amplified (*vast*), or problems dramatised (*devastating*). Such language has been termed “hype” and defined as “hyperbolic or subjective language employed to glamorise, promote, or exaggerate aspects of research” [1].

Our work is motivated by concern about increasing levels of hype in biomedical communication. Promotional language has risen sharply across the National Institutes of Health (NIH) funding ecosystem, from calls for grants, to applications, and subsequent publications [2, 3, 4]. Comparable trends are seen in research articles in other fields, press releases, and media reports [5, 6]. Promotional words such as *groundbreaking* or *transformative* are rarely justified and risk undermining objective assessment, thereby impeding the development of further studies, policies, clinical practice, and knowledge translation [7].

The automatic identification of hype poses challenges for an NLP approach. Firstly, promotionality is often context-dependent. While some terms are used almost invariably with a promotional connotation, others may also carry technical or neutral meanings. Secondly, judgments about whether a term is promotional are inherently subjective, making it difficult to establish ground truth for training and evaluation. Finally, because hype should be seen not as a binary distinction but a continuum of intensity, with terms conveying different levels of promotion, a measure of intensity of individual terms is needed.

In this paper, we obtain promotional intensity ratings for 303 adjectives common in biomedical texts and carrying a promotional meaning. The selection of terms is based on prior analyses of promotional language in NIH grant abstracts [2]. We apply Best-Worst Scaling (BWS) to overcome limitations of traditional rating scales, and show that raters provide high levels of agreement across annotations (split-half reliability = 0.87). Our main contribution is the **Hyplex** resource, a lexicon of 303 hype adjectives in biomedical writing with human-derived promotional intensity scores.

2 Related Work

2.1 Promotional Language

In linguistics, promotion is usually discussed as part of the broader study of evaluative meaning, with several proposed frameworks [8, 9, 10]. All describe systems or resources that enable writers to take positions and attribute value in texts, e.g., markers of attitude (*important, valuable*), epistemic stance (*likely, possibly*), degree (*highly effective* vs. *somewhat effective*), and reader alignment (*notably, as we know*). Promotion can be understood as the use of such resources with the intent of encouraging a favourable evaluation. Linguistic analyses tend to rely on

manual interpretation to determine if and how expressions convey evaluation, due to dependency on context, while NLP approaches have typically operationalised evaluation as sentiment (positive, neutral, or negative). Recent NLP work [11] has addressed promotional language directly by formalising hype detection in NIH grant texts as a binary classification problem.

2.2 Related Lexicons

Work in sentiment analysis has produced several general-purpose resources for modeling subjectivity and evaluation [12, 13], through human annotation and semi-automatic induction algorithms. Affective lexicons with interval-scaled scores have been constructed using Best-Worst Scaling (BWS) [14, 15]; these lexicons are both built for broad-domain text and are not specific to scientific or promotional language. In the biomedical field, Millar et al. [2] identified 139 adjectives that often carry promotional meaning through corpus analyses of successful NIH grant application abstracts, showing sharp increases in frequency over recent decades. The lexicon has also been used to show that the proportion of promotional language in proposals is strongly associated with funding success, innovativeness, and subsequent citation impact [16, 17]. However, this resource is limited to 139 adjectives and does not quantify promotional intensity (e.g., *novel* < *revolutionary*).

2.3 Best-Worst Scaling

Best-Worst Scaling (BWS) [18] is a comparative judgment method in which respondents are shown a small set of items and asked to select the one that best and the one that least matches a target property. BWS reduces biases common in rating scales and produces more consistent and discriminating results across individuals [19, 20]. Each judgment implicitly generates a series of pairwise preference statements. For example, when presented with four items, if A is judged best and D worst, this implies $A > B, A > C, A > D, B > D$, and $C > D$. When aggregated across many judgments and participants, these comparative data can produce stable estimates of each item' s relative position on an underlying scale [21]. BWS has been applied in fields such as marketing, psychology, and linguistics [22, 23], and has been used to create intensity lexicons for affective variables [14, 15].

Category	Examples	#
ATTITUDE	outstanding, impressive	20
IMPORTANCE	critical, essential	29
NOVELTY	novel, groundbreaking	31
PROBLEM	devastating, stark	30
QUALITIES	talented, cohesive	44
RIGOUR	systematic, robust	44
SCALE	large-scale, extensive	48
UTILITY	effective, transformative	57
	Total	**303**

Table 1: Sematic categories, example adjectives and number of items in HYPLEX (minus low and neutral anchors).

3 Methods

3.1 Selection of terms

We limit our focus to adjectives, as this class is most closely associated with evaluation [9]. Seed terms were drawn from [2], which identified 139 "hype" adjectives through longitudinal analysis of a corpus of 901,717 abstracts from successful NIH funding applications. To expand coverage, we used WordNet [24] and ChatGPT[1)] to generate near-synonyms for each seed adjective. Generated terms were checked against the NIH corpus, a collection of 901,717 abstracts from successful NIH grant applications spanning 1985-2020 [2], and were retained if they occurred more than ten times and conveyed promotional meaning. The combined final list contained 303 adjectives, distributed across categories as shown in Table 1.

3.2 Annotation via Best-Worst Scaling

Participants. Ten annotators took part: six postgraduate students, two researchers, and two university Professors fluent in English, including four native speakers. All participants provided informed consent prior to participation. Student annotators were compensated for their time at rates consistent to local minimum wage standards, while researchers and Professors generously dedicated their free time to assist us.

Design. Annotation followed a Balanced Incomplete Block Design (BIBD) [25] that was constructed separately for each semantic category. Within each BIBD, adjectives were presented in sets of four, and annotators selected the most promotional ("best") and least promotional ("worst")

1) https://openai.com/index/introducing-gpt-5/

Anchor Type	Examples	Interpretation
High (≈ 1.00)	revolutionary, flawless	Defines upper bound for category.
Neutral (> Low)	standard, adequate	Quality-check; not used in scaling.
Low (≈ 0.00)	unoriginal, unreliable	Defines lower bound for category.

Table 2: Anchor types, examples, and interpretation. Anchors relevant to each category were embedded within the BIBDs to provide consistent reference points for cross-BIBD scale calibration.

term. Anchors relevant to the category were embedded within the BIBDs to provide consistent reference points for cross-BIBD scale calibration. High anchors defined the upper bound of promotional intensity, low anchors defined the lower bound, and neutral anchors served as quality-check items expected to fall directly above the lower bound. Table 2 summarizes the role and interpretation of the anchors with examples for NOVELTY and RIGOUR.

Implementation. BIBDs were created using the `bwsTools` package in R [26] and divided into surveys by semantic category, administered over five separate days using a custom BWS platform. Each survey began with practice items (unrelated to promotional language), an introduction to promotional language in science, and a category-specific explanation. Attention checks (minimum two, maximum four per survey) were inserted to ensure participants were not responding mechanically.

Scoring and aggregation. The final promotional intensity scores were calculated using a difference-scoring procedure [27]. For each participant, the n_- times an adjective was selected as "least promotional" was subtracted from the n_+ times it was selected as "most promotional", and the resulting difference was standardized by the total n times the item appeared.

Within each semantic category, participant-level difference scores were rescaled to a $[0, 1]$ range, producing participant-level intensities for each BIBD. These scores were then linearly adjusted using the observed mean of the embedded high (1) and low anchors (0) as reference points. Anchor-based calibration preserved within-category rank order while approximately aligning category-specific BIBDs to a comparable intensity range. The calibrated participant-level scores were averaged across annotators to obtain the final intensity estimate for each adjective [27].

Category	Low	Neutral	High
ATTITUDE	0.05	0.09	0.84
IMPORTANCE	0.00	0.11	0.86
NOVELTY	0.07	0.09	0.88
PROBLEM	0.06	0.11	0.93
QUALITIES	0.03	0.16	0.82
RIGOUR	0.05	0.20	0.92
SCALE	0.06	0.14	0.78
UTILITY	0.10	0.17	0.88

Table 3: Mean intensity scores (0-1) for low, neutral, and high anchors by semantic category.

Category	SHR	95% CI
ATTITUDE	0.94	[0.91, 0.97]
NOVELTY	0.92	[0.87, 0.96]
PROBLEM	0.92	[0.87, 0.96]
QUALITIES	0.89	[0.84, 0.93]
SCALE	0.88	[0.81, 0.93]
IMPORTANCE	0.87	[0.81, 0.92]
UTILITY	0.78	[0.71, 0.85]
RIGOUR	0.75	[0.63, 0.85]

Table 4: Split-half reliability (SHR) (mean r) of promotional intensity judgments by semantic category.

Internal reliability. Split-half reliability (SHR) was used to assess the internal consistency of the annotations. For each semantic category, annotators were randomly divided into two groups, and mean adjective scores were independently computed for each half using the scaled difference scores. The correlation between the two halves estimates the reliability of item ranking within each category. The process was repeated 500 times to obtain mean and confidence interval estimates. The same procedure was also applied across all items to evaluate reliability at the full-lexicon level.

Anchor validation. To confirm that the anchors operated as intended, we examined the mean intensity scores for each semantic category to verify that (i) low anchors consistently received the lowest scores, (ii) high anchors consistently received the highest scores, and (iii) neutral anchors were positioned just above the low anchors but below the other items.

4 Results

Data Overview. The ten annotators completed the eight category-specific surveys, generating a total of 4,340 best-worst judgments. All attention-check questions were answered correctly, suggesting adequate engagement with

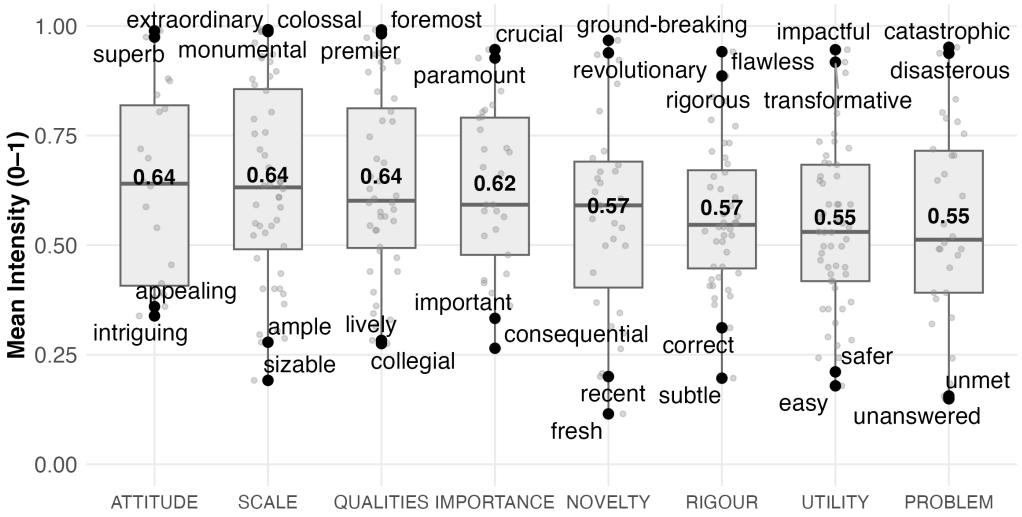

Figure 1: Distribution of Promotional Intensity Scores Across HYPLEX Categories (mean intensity scores shown on boxplots).

the task. Following the procedure outlined in Section 3.2, we obtained the HYPLEX, a lexicon of 303 adjectives with category-specific intensity scores.

Anchor validation. Table 3 summarizes the mean intensity scores of the low, neutral, and high anchors across categories before calibration. In every category, the anchors appeared in the expected order (low < neutral < high). Low and neutral anchors consistently ranked as the least promotional items, while high anchors tended to appear at or near the top of their respective distributions. Thus, the anchors generally functioned as intended and were used as reference points for the calibration of BIBDs within individual semantic categories.

Reliability. As shown in Table 4, split-half reliability (SHR) across categories ranged from r = 0.75 (RIGOUR) to r = 0.94 (ATTITUDE). The more reliable categories (e.g., ATTITUDE, NOVELTY) contain adjectives with clearer evaluative meanings (e.g. *excellent*, *groundbreaking*), whereas UTILITY and RIGOUR involve more technical, context-dependent terms (e.g., *effective*, *accurate*, *precise*) which are likely to be interpreted differently by individual annotators. The overall SHR across all adjectives was $r = 0.87, 95\%$ CI $[0.85, 0.89]$, consistent with accepted thresholds for good internal consistency for psychometric scales (≥ 0.80) [28].

Distribution of promotional intensity scores. Figure 1 shows the distribution of promotional intensity scores across categories and the two most and least promotional adjectives in each category. Excluding low and neutral anchors, the mean of promotional intensity scores ranged from 0.55 (PROBLEM) to 0.64 (SCALE). Categories showed broadly similar distributions, with interquartile ranges spanning roughly 0.35 – 0.80. Categories SCALE, QUALITIES, and ATTITUDE included a slightly higher proportion of strongly promotional terms. NOVELTY showed the widest spread of scores (0.12 – 0.97). Furthermore, in all cases, these extremes align with intuitive expectations, e.g., *catastrophic* and *disastrous* rank top in PROBLEM, while *unmet* and *unanswered* rank bottom. We refer the reader to Table 5 in the appendix for the full ranking of adjectives.

5 Conclusions

Building on prior analyses of NIH funding applications, we developed the HYPLEX resource - an intensity-scaled lexicon of 303 promotional adjectives attested in biomedical research writing. The lexicon showed overall internal consistency within preferred thresholds for psychometric scale quality and produced intuitively ordered rankings (e.g., *important* < *vital* < *critical* < *paramount*).

We see many potential applications and extensions, such as supplying interpretable features for hype-detection systems and for developing hype-aware word and sentence embeddings, serving as a gold-standard reference for evaluating automatic methods of determining promotionality, and examining whether levels of promotionality influence readers' evaluation of research, or measurable outcomes such as funding success, publication, and citation impact.

Acknowledgement

This study was supported by grant No. 25K00851 from the Japan Society for the Promotion of Science. Additionally, this paper is based on results obtained from project JPNP25006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

[1] Neil Millar, Françoise Salager-Meyer, and Brian Budgell. "it is important to reinforce the importance of⋯" : 'hype' in reports of randomized controlled trials. **English for Specific Purposes**, Vol. 54, pp. 139–151, 2019.

[2] Neil Millar, Bojan Batalo, and Brian Budgell. Trends in the use of promotional language (hype) in abstracts of successful national institutes of health grant applications, 1985-2020. **JAMA network open**, Vol. 5, No. 8, pp. e2228676–e2228676, 2022.

[3] Neil Millar, Bojan Batalo, and Brian Budgell. Trends in the use of promotional language (hype) in national institutes of health funding opportunity announcements, 1992-2020. **JAMA Network Open**, Vol. 5, No. 11, pp. e2243221–e2243221, 2022.

[4] Neil Millar, Bojan Batalo, and Brian Budgell. Promotional language (hype) in abstracts of publications of national institutes of health–funded research, 1985-2020. **JAMA Network Open**, Vol. 6, No. 12, pp. e2348706–e2348706, 2023.

[5] Petroc Sumner, Solveiga Vivian-Griffiths, Jacky Boivin, Andy Williams, Christos A Venetis, Aimée Davies, Jack Ogden, Leanne Whelan, Bethan Hughes, Bethan Dalton, et al. The association between exaggeration in health related science news and academic press releases: retrospective observational study. **Bmj**, Vol. 349, , 2014.

[6] Nils B Weidmann, Sabine Otto, and Lukas Kawerau. The use of positive words in political science language. **PS: Political Science & Politics**, Vol. 51, No. 3, pp. 625–628, 2018.

[7] Howard Bauchner and Frederick P Rivara. The scientific communication ecosystem: the responsibility of investigators. **The Lancet**, Vol. 400, No. 10360, pp. 1289–1290, 2022.

[8] Susan Hunston and Geoffrey Thompson. **Evaluation in text: Authorial stance and the construction of discourse: Authorial stance and the construction of discourse**. Oxford University Press, UK, 2000.

[9] James R Martin and Peter R White. **The language of evaluation**, Vol. 2. Springer, 2003.

[10] Ken Hyland. Metadiscourse: Exploring interaction in writing. **Continuum**, 2005.

[11] Bojan Batalo, Erica K. Shimomoto, and Neil Millar. Hype or not? formalizing automatic promotional language detection in biomedical research, 2025.

[12] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis, 2009.

[13] Song Feng, Jun Seok Kang, Polina Kuznetsova, and Yejin Choi. Connotation lexicon: A dash of sentiment beneath the surface meaning, 2013.

[14] Saif Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words, July 2018.

[15] Saif Mohammad and Felipe Bravo-Marquez. Emotion intensities in tweets, August 2017.

[16] Hao Peng, Huilian Sophie Qiu, Henrik Barslund Fosse, and Brian Uzzi. Promotional language and the adoption of innovative ideas in science. **Proceedings of the National Academy of Sciences**, Vol. 121, No. 25, p. e2320066121, 2024.

[17] Huilian Sophie Qiu, Hao Peng, Henrik Barslund Fosse, Teresa K Woodruff, and Brian Uzzi. Use of promotional language in grant applications and grant success. **JAMA network open**, Vol. 7, No. 12, pp. e2448696–e2448696, 2024.

[18] Jordan J Louviere and George G Woodworth. Best-worst scaling: A model for the largest difference judgments. Technical report, working paper, 1991.

[19] Adam Finn and Jordan J Louviere. Determining the appropriate response to evidence of public concern: the case of food safety. **Journal of Public Policy & Marketing**, Vol. 11, No. 2, pp. 12–25, 1992.

[20] Terry N Flynn and Anthony AJ Marley. Best-worst scaling: theory and methods. In **Handbook of choice modelling**, pp. 178–201. Edward Elgar Publishing, 2014.

[21] Jordan J. Louviere, Terry N. Flynn, and A. A. J. Marley. **Best-Worst Scaling: Theory, Methods and Applications**. Cambridge University Press, 2015.

[22] Christopher Crocker and David MH Thomson. Anchored scaling in best–worst experiments: A process for facilitating comparison of conceptual profiles. **Food Quality and Preference**, Vol. 33, pp. 37–53, 2014.

[23] Svetlana Kiritchenko and Saif Mohammad. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling, 2016.

[24] George A Miller. Wordnet: a lexical database for english, 1995.

[25] William G Cochran and Gertrude M Cox. **Experimental designs**. John Wiley & Sons, 1957.

[26] Mark H White II. bwstools: An r package for case 1 best-worst scaling. **Journal of choice modelling**, Vol. 39, p. 100289, 2021.

[27] Jordan Louviere, Ian Lings, Towhidul Islam, Siegfried Gudergan, and Terry Flynn. An introduction to the application of (case 1) best–worst scaling in marketing research. **International journal of research in marketing**, Vol. 30, No. 3, pp. 292–303, 2013.

[28] Godfred O Boateng, Torsten B Neilands, Edward A Frongillo, Hugo R Melgar-Quiñonez, and Sera L Young. Best practices for developing and validating scales for health, social, and behavioral research: a primer. **Frontiers in public health**, Vol. 6, p. 149, 2018.

ATTITUDE	IMPORTANCE	NOVELTY	PROBLEM	QUALITIES	RIGOUR	SCALE	UTILITY
1 extraordinary	1 crucial	1 ground-breaking	1 catastrophic	1 foremost	1 flawless	1 colossal	1 impactful
2 superb	2 paramount	2 revolutionary	2 disasterous	2 premier	2 rigorous	2 monumental	2 transformative
3 incredible	3 critical	3 paradigm-shifting	3 devastating	3 leading	3 meticulous	3 staggering	3 high-performance
4 phenomenal	4 indispensable	4 unprecedented	4 ruinous	4 brilliant	4 robust	4 enormous	4 high-yielding
5 astonishing	5 pivotal	5 unheard-of	5 hopeless	5 prestigious	5 high-level	5 innumerable	5 perfect
6 outstanding	6 imperative	6 unparalleled	6 deadly	6 distinguished	6 high-standard	6 mammoth	6 ideal
7 fascinating	7 vital	7 one-of-a-kind	7 dire	7 exceptional	7 sophisticated	7 overpowering	7 optimal
8 thrilling	8 ultimate	8 unequaled	8 grave	8 stellar	8 painstaking	8 gigantic	8 synergistic
9 impressive	9 momentous	9 unique	9 miserable	9 pre-eminent	9 precise	9 tremendous	9 high-performing
10 exciting	10 essential	10 trailblazing	10 perilous	10 veteran	10 methodical	10 massive	10 meaningful
11 remarkable	11 invaluable	11 unrivaled	11 grim	11 renowned	11 exacting	11 overwhelming	11 valuable
12 excellent	12 fundamental	12 never-before-seen	12 bleak	12 accomplished	12 thorough	12 greatest	12 effective
13 rewarding	13 urgent	13 pioneering	13 desperate	13 thriving	13 accurate	13 vast	13 durable
14 surprising	14 foundational	14 game-changing	14 shocking	14 talented	14 scientific	14 immense	14 efficient
15 attractive	15 pressing	15 cutting-edge	15 dismal	15 gifted	15 error-free	15 sweeping	15 scalable
16 confident	16 profound	16 inventive	16 frightening	16 seasoned	16 powerful	16 worldwide	16 seamless
17 notable	17 high-priority	17 incomparable	17 alarming	17 forward-thinking	17 fine-grained	17 myriad	17 opportune
18 interesting	18 prime	18 innovative	18 daunting	18 reputable	18 strict	18 countless	18 efficacious
19 appealing	19 decisive	19 radical	19 stark	19 longstanding	19 detailed	19 exhaustive	19 concrete
20 intriguing	20 significant	20 latest	20 formidable	20 established	20 careful	20 comprehensive	20 expandable
	21 key	21 emerging	21 intimidating	21 dedicated	21 disciplined	21 global	21 productive
	22 necessary	22 novel	22 scarce	22 credentialed	22 advanced	22 largest	22 sustainable
	23 major	23 first	23 disturbing	23 certified	23 discriminating	23 transdisciplinary	23 useful
	24 chief	24 up-to-date	24 serious	24 successful	24 strong	24 biggest	24 purposeful
	25 compelling	25 original	25 dramatic	25 dynamic	25 elegant	25 far-reaching	25 streamlined
	26 strategic	26 creative	26 worrying	26 skilled	26 structured	26 fastest	26 extensible
	27 influential	27 newest	27 troubling	27 knowledgeable	27 exact	27 generous	27 tailored
	28 important	28 imaginative	28 elusive	28 qualified	28 systematic	28 multifarious	28 intuitive
	29 consequential	29 up-and-coming	29 unmet	29 senior	29 quality	29 huge	29 fruitful
		30 recent	30 unanswered	30 experienced	30 coordinated	30 top	30 dependable
		31 fresh		31 rising	31 reproducible	31 diverse	31 tactical
				32 vibrant	32 cohesive	32 intense	32 straightforward
				33 ambitious	33 verifiable	33 extensive	33 constructive
				34 promising	34 complex	34 expansive	34 accessible
				35 energetic	35 integrated	35 international	35 adaptable
				36 holistic	36 refined	36 abundant	36 actionable
				37 intellectual	37 logical	37 complete	37 implementable
				38 supportive	38 unified	38 considerable	38 practical
				39 motivated	39 empirical	39 deeper	39 generalizable
				40 integrative	40 repeatable	40 immediate	40 practicable
				41 committed	41 nuanced	41 wide-ranging	41 tangible
				42 cohesive	42 organized	42 substantial	42 transferable
				43 lively	43 correct	43 broad	43 maintainable
				44 collegial	44 subtle	44 plenty	44 well-timed
						45 prompt	45 timely
						46 instant	46 rich
						47 ample	47 easy-to-use
						48 sizable	48 user-friendly
							49 deployable
							50 self-explanatory
							51 usable
							52 viable
							53 ready
							54 simple
							55 economical
							56 safer
							57 easy

Table 5: Category-wise ranking of adjectives in the HYPLEX lexicon, ordered by mean promotional intensity score.