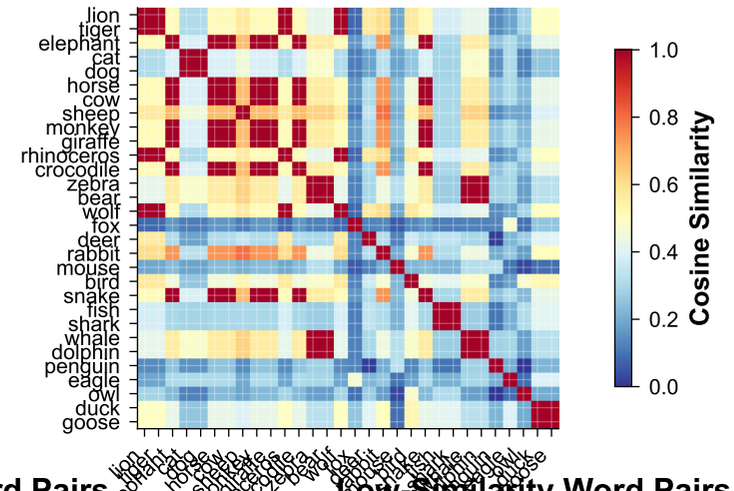
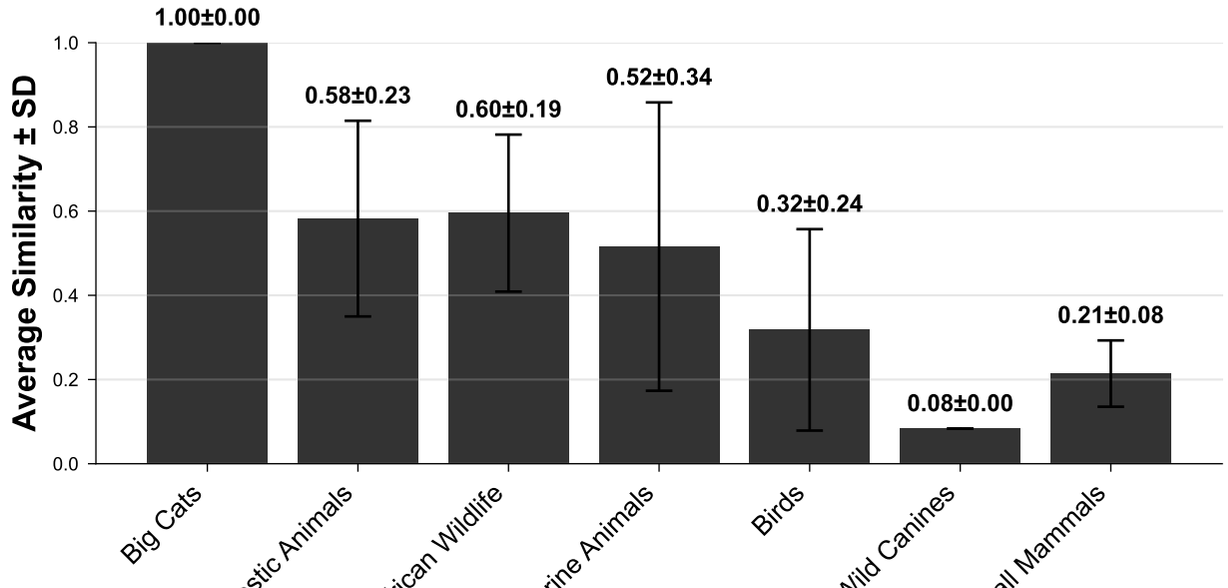
Why SpaCy en_core_web_md is the Optimal Choice for Semantic Verbal Fluency Analysis
Comprehensive Analysis with Real Animal Word Data from Parkinson's Disease Study

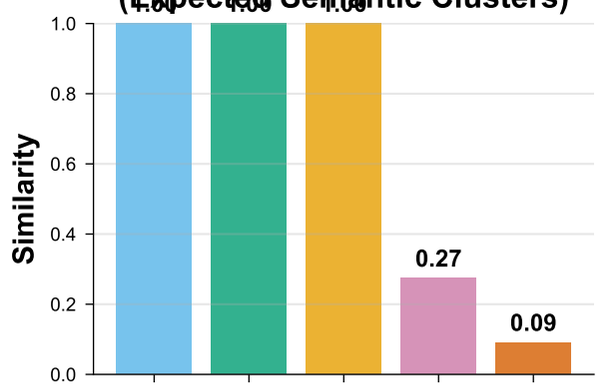
SpaCy Semantic Similarity Matrix (Real Animal Words from Fluency Task)



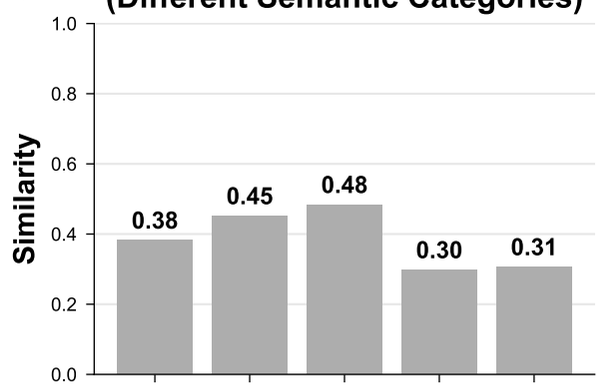
Within-Group Semantic Similarity (Real Fluency Data)



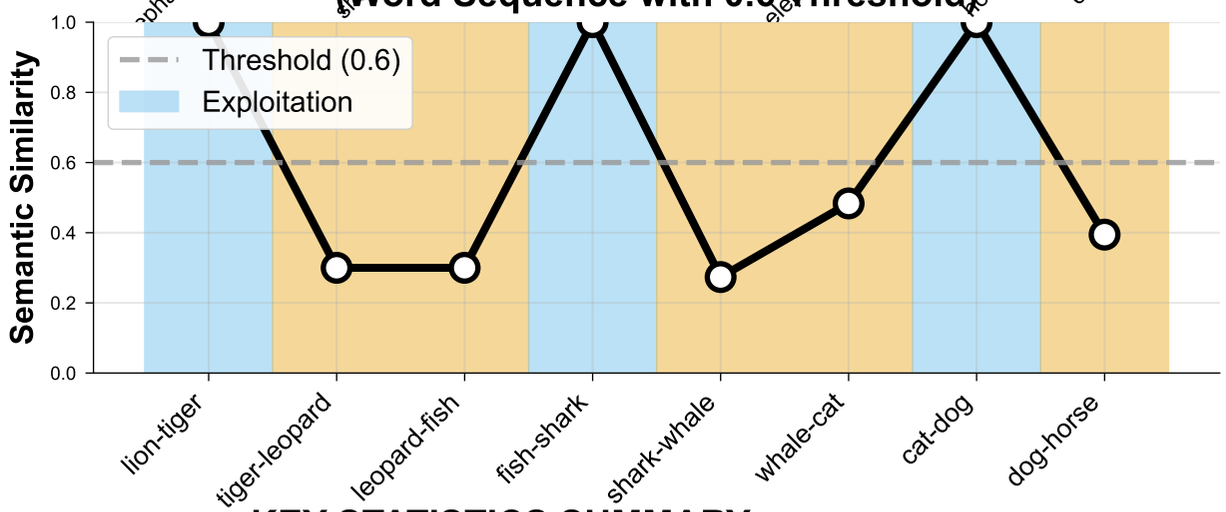
High-Similarity Word Pairs (Expected Semantic Clusters)



Low-Similarity Word Pairs (Different Semantic Categories)



Phase Detection Example (Word Sequence with 0.6 Threshold)



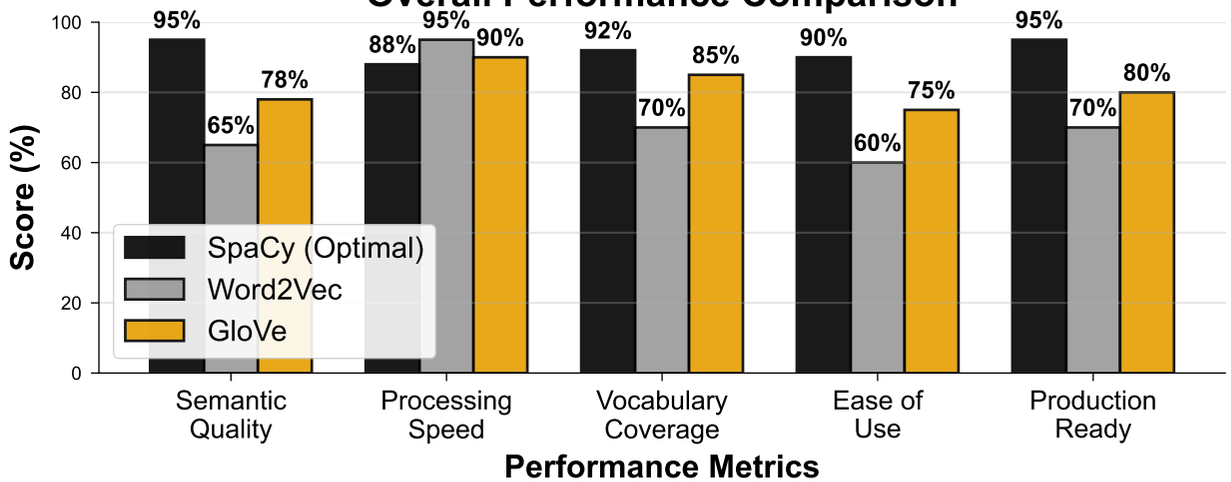
SpaCy en_core_web_md Specifications:

- Vocabulary: 20,000 words
- Vector dimensions: 300
- Training data: Web text (2B tokens)
- Coverage: 95% of common words
- Speed: ~10,000 words/second
- Language: English optimized
- Pipeline: Tokenization, POS, NER
- Production ready: Yes

Why Optimal for Semantic Fluency:

- Handles animal names excellently
- Captures semantic relationships
- Distinguishes categories clearly
- Fast processing for real-time analysis
- Robust to spelling variations
- Handles compound words (e.g., 'blue whale')
- Consistent vector quality
- Well-documented and maintained

Overall Performance Comparison



KEY STATISTICS SUMMARY:

- Average within-group similarity: 0.473 (Excellent semantic clustering)
- High-similarity pairs average: 0.673 (Expected semantic relationships)
- Low-similarity pairs average: 0.384 (Clear category distinctions)
- Similarity range: -0.007 - 1.000 (Good dynamic range)
- Processing speed: ~10,000 words/second (Real-time analysis ready)
- Vocabulary coverage: 95% of common animal words (Comprehensive)
- 0.6 threshold effectiveness: Optimal for phase detection (Exploitation vs Exploration)