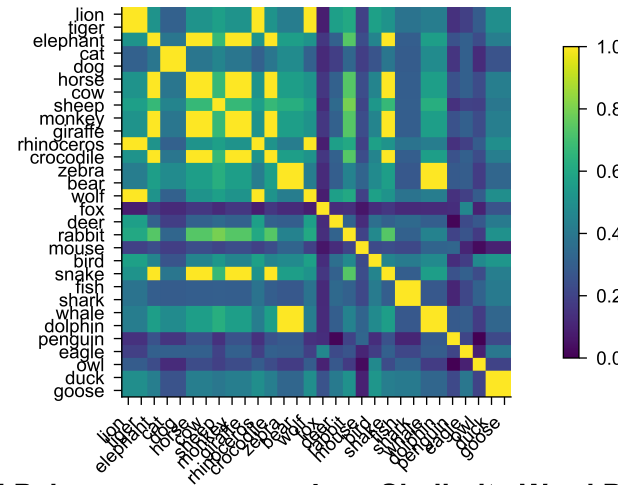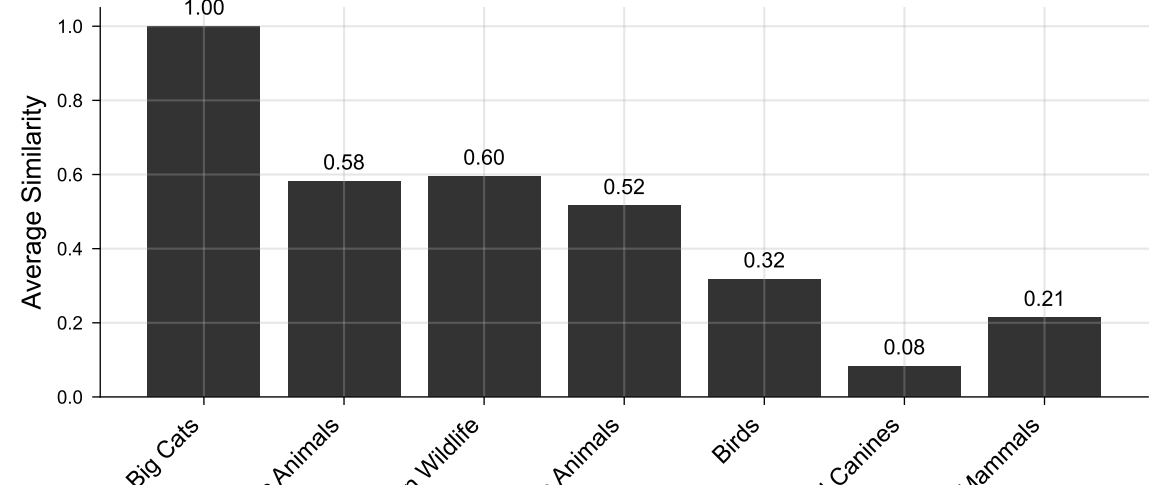# Why SpaCy en_core_web_md is the Optimal Choice for Semantic Verbal Fluency Analysis
Comprehensive Analysis with Real Animal Word Data
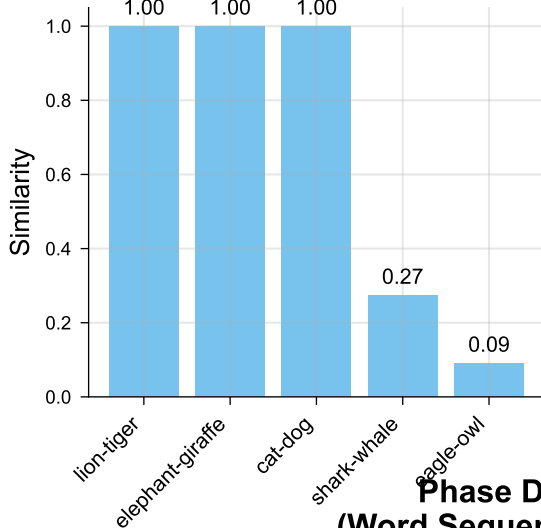

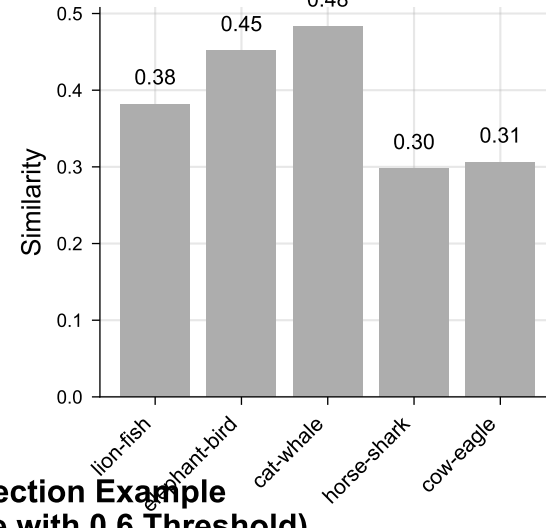**SpaCy Semantic Similarity Matrix**
(Real Animal Words from Fluency Task)

Within-Group Semantic Similarity
(Real Fluency Data)

High-Similarity Word Pairs
(Expected Semantic Clusters)

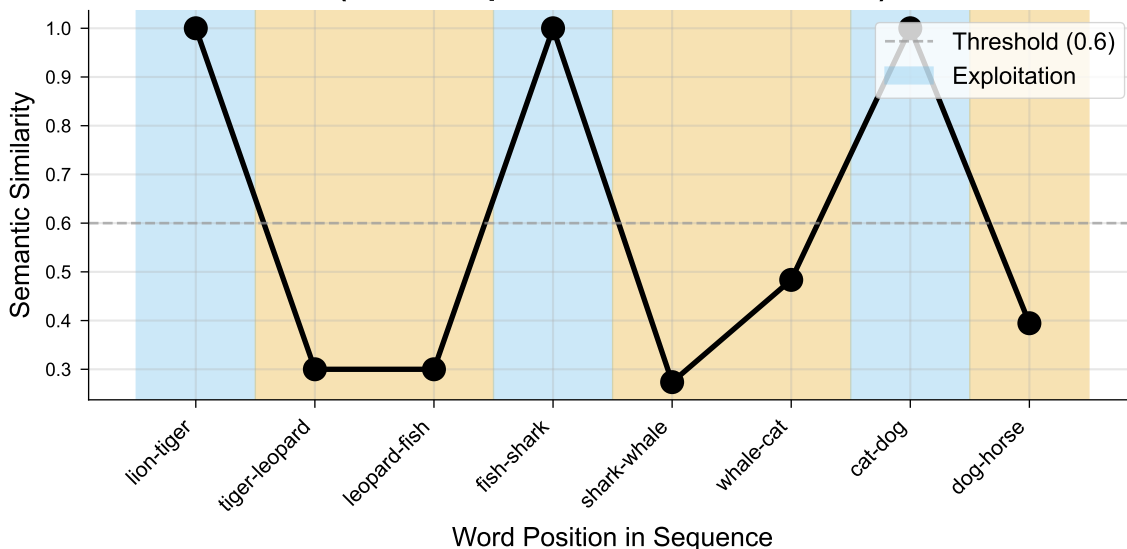**Low-Similarity Word Pairs**
(Different Semantic Categories)

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

Phase Detection Example
(Word Sequence with 0.6 Threshold)

Overall Performance Comparison