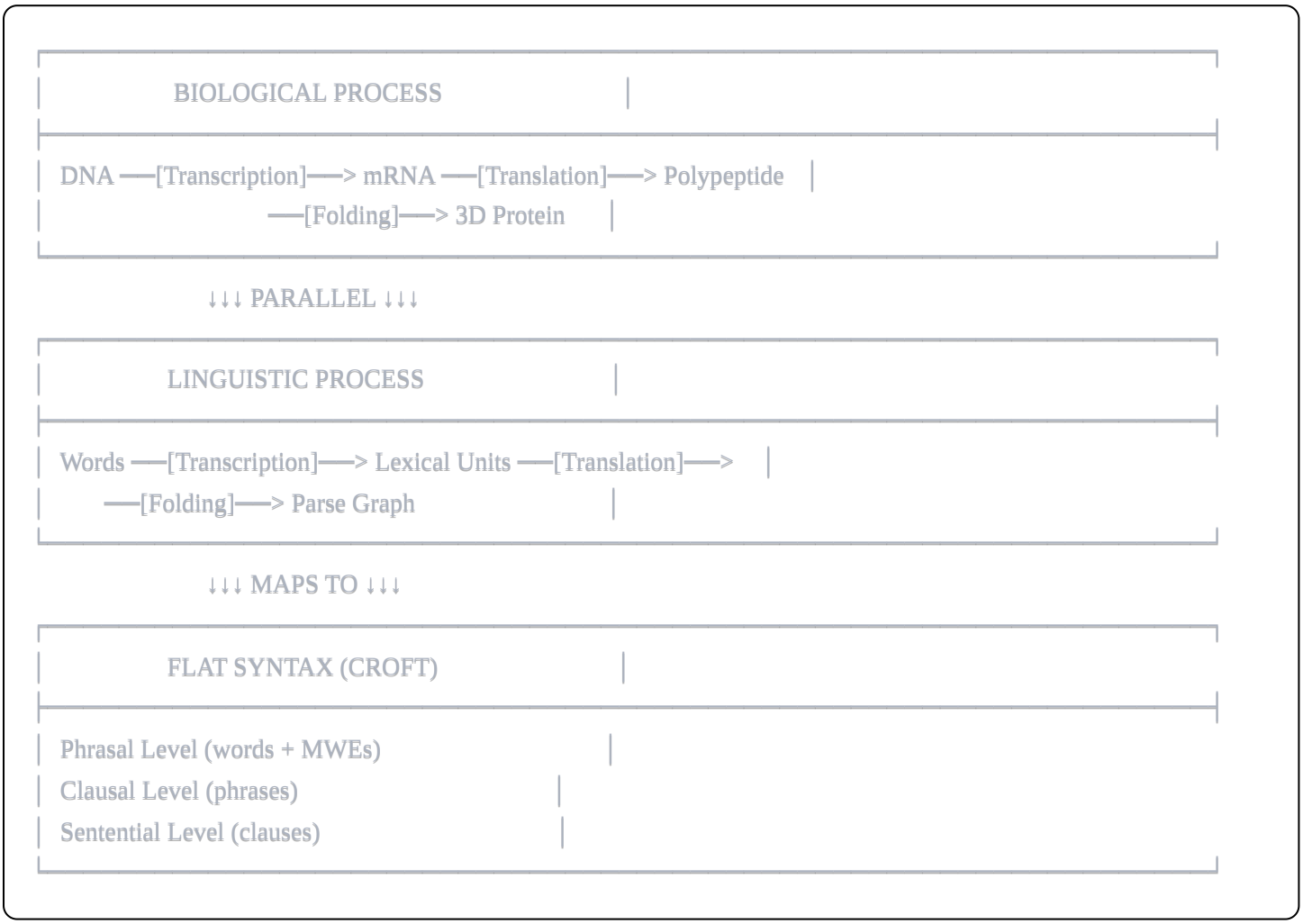
Visual Guide: Three-Stage Parsing Framework

Biological-Linguistic Parallel Reference

QUICK REFERENCE: Three-Stage Mapping



STAGE 1: TRANSCRIPTION (Lexical Level)

Biological Analogy

```
| Ready for translation |
└──────────────────────┘
```

Linguistic Processing

```
Word Sequence → Lexical Units with Types + Features

Input:   "tomei café da manhã"

Process:
1. Extract UD features for each word
  tomei:  VerbForm=Fin, Mood=Ind, Tense=Past → type V
  café:   Gender=Masc, Number=Sing → type E
  da:     [function word] → type F
  manhã:  Gender=Fem, Number=Sing → type E

2. Detect MWE: "café da manhã" (breakfast)
  Activate prefix hierarchy:
  - [café]        : 1/3 activation
  - [café da]     : 2/3 activation
  - [café da manhã]: 3/3 activation ✓ STABLE → Aggregate

Output:
  [tomei: V] [café_da_manhã: E]
  (2 stable lexical units ready for phrase building)
```

Croft's Phrasal Level

```
Flat Syntax Annotation:
"tomei + café^da^manhã"

Labels:
tomei:        Head (phrasal) → Pred (clausal) → Main (sentential)
café^da^manhã: Head (phrasal) → Arg (clausal) → Main (sentential)

Boundary markers:
- Space: separates words
- ^: joins MWE components (café^da^manhã treated as unit)
- +: separates phrasal constituents (comes later at Translation)
```

Feature as Chemical Properties

```
Morphological Feature ← → Amino Acid Property

NOUN Features:
```

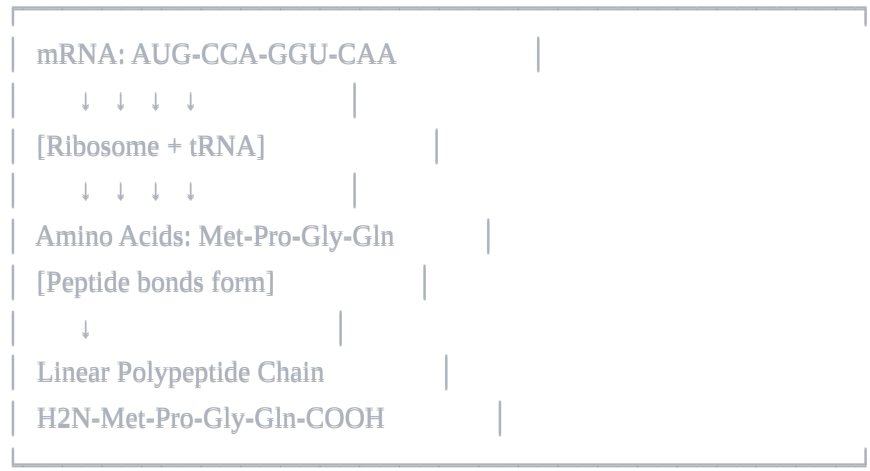
Gender=Masc ← → Hydrophobic (clusters with same)
Number=Sing ← → Small size (fits specific slots)
Case=Nom ← → Positively charged (attracts to Subject)

VERB Features:
VerbForm=Fin ← → Catalytic active site (can be predicate)
Mood=Ind ← → Neutral pH (default conditions)
Tense=Past ← → Temporal state marker

STAGE 2: TRANSLATION (Phrasal Level)

Biological Analogy

Ribosomal Translation:



Linguistic Processing

Lexical Units → Phrasal Constituents

Input: [tomei: V] [café_da_manhã: E] [cedo: ADV]

Process:
1. Build phrases around lexical heads

tomei (V) predicts:
- Subject: [implicit pro-drop in Portuguese]
- Object: expects entity (E type)

café_da_manhã (E) matches Object prediction:
- Type: E ✓
- Features compatible ✓
- Link: tomei ──[OBJ]──> café_da_manhã

cedo (ADV):
- Modifies predicate
- Link: tomei ──[ADV]──> cedo

2. Identify phrase types (Croft's labels):
[Pred: tomei]
[Arg: café_da_manhã]
[FPM: cedo]

Output:
Three separate phrases, ready for sentence integration

Feature-Driven Linking

Spanish Example: "la casa grande"

Lexical units with features:

```
┌──────────────────────────────────────────────┐
│ la:     DET  Gender=Fem, Number=Sing, Definite=Def  │
│ casa:   NOUN Gender=Fem, Number=Sing                │
│ grande: ADJ  Number=Sing                            │
└──────────────────────────────────────────────┘
```

Agreement checking (like peptide bonding):

```
┌──────────────────────────────────────────────┐
│ la → casa:                                   │
│   Gender: Fem = Fem ✓ [H-bond 1]             │
│   Number: Sing = Sing ✓ [H-bond 2]           │
│   Link strength: HIGH                        │
│                                              │
│ grande → casa:                               │
│   Number: Sing = Sing ✓ [H-bond 3]           │
│   Link strength: MEDIUM                      │
└──────────────────────────────────────────────┘
```

Result: [NP: la + casa + grande]
Unified features: Gender=Fem, Number=Sing, Definite=Def

Croft's Clausal Level

Flat Syntax Annotation:
"tomei + café^da^manhã + cedo ."

Clausal CE Labels:

```
tomei:          Pred (predicate - main verb)
café^da^manhã: Arg  (argument - object of verb)
cedo:          FPM  (flagged phrase modifier - adverb)

Boundary markers:
- +: separates clausal constituents (phrases)
```

Chemical Bonding Analogy

```
Peptide Bond Formation ← → Phrase Construction

Amino Acids:          Words:
 Met   Pro            tomei  café_da_manhã
  |    |              |      |
  [peptide bond]       [dependency link]
  |                    |
 Met-Pro               tomei → café_da_manhã

Bond strength depends on:
- Local properties      - Feature compatibility
- pH, temperature       - Context, word order
- Side chain chemistry  - Morphological features
```

STAGE 3: FOLDING (Sentential Level)

Biological Analogy

```
Protein Folding:

┌──────────────────────────────────────────────────────┐
│  Linear Polypeptide                                  │
│  H2N-Met-Ala-Gly-...-Leu-Pro-COOH                    │
│              ↓                                       │
│  [Hydrophobic collapse]                              │
│  [Secondary structures form]                         │
│  [Disulfide bridges connect distant Cys]             │
│              ↓                                       │
│  3D Folded Protein                                   │
│        ____                                          │
│       /    \   S-S bridge                            │
│      | ●●● |  /                                      │
│       \____/  /                                      │
│         |    /                                       │
```

```
│        [core] [surface]                  │
└──────────────────────────────────────────┘
```

Linguistic Processing

Phrases → Complete Parse Graph

Example: "O menino que eu vi chegou cedo"
(The boy that I saw arrived early)

Phrases from Translation:

```
┌──────────────────────────────────────────┐
│ [Arg: o menino]                          │
│ [Rel: que eu vi]                         │
│ [Pred: chegou]                           │
│ [FPM: cedo]                              │
└──────────────────────────────────────────┘
```

Folding operations:
1. Identify main clause backbone:
   chegou (root predicate)
   o menino (subject)

2. Attach relative clause (LONG-DISTANCE):
   que → menino (relative pronoun to antecedent)
   que ← vi (as object of embedded verb)

   THIS CREATES CROSSING EDGE (non-projective)
   Like disulfide bridge in proteins!

3. Attach adverb:
   cedo → chegou

Final structure:

```
        chegou (ROOT)
        /    \
   menino    cedo
      |
     que ←──────┐
    /   \       │ [crossing edge!]
  eu    vi ─────┘
```

Linear: o_1 $menino_2$ que_3 eu_4 vi_5 $chegou_6$ $cedo_7$
Graph: $menino_2$ ←─ $chegou_6$ (crosses 3,4,5)

Croft's Sentential Level

Flat Syntax Annotation:
"O menino {que + eu + vi} + chegou + cedo ."

Sentential CE Labels:
Main clause: [o menino] [chegou] [cedo]
Relative clause: {que eu vi}

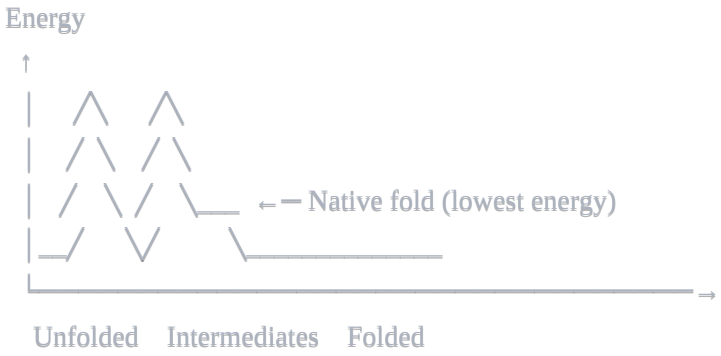
Boundary markers:
- { }: marks interruption (center-embedding)
- #: would separate multiple clauses (not needed here)
- .: sentence boundary

Interpretation:
"O menino {que + eu + vi} + chegou + cedo"
 Main:Arg Rel:clause Main:Pred FPM

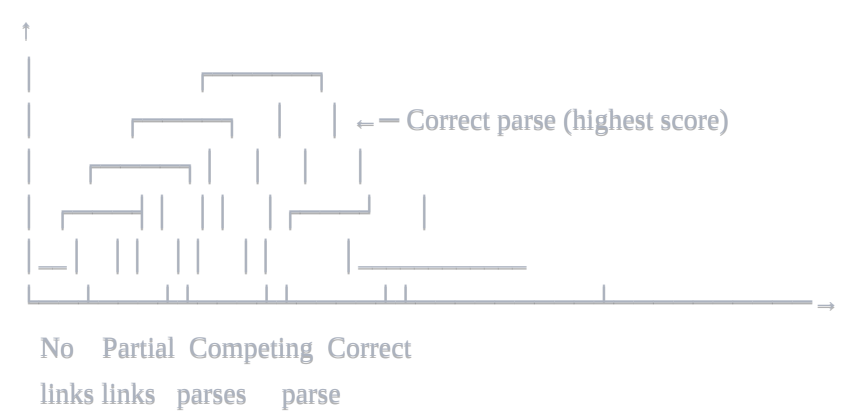
The relative clause interrupts the main clause's
subject NP, just like an insertion sequence in DNA.

Energy Landscape Analogy

Protein Energy Landscape:



Parse Score Landscape:



Both systems:
- Multiple local minima/maxima (ambiguity)
- Global optimum (correct structure)
- Energy barriers (activation thresholds)
- Quality control (garbage collection)

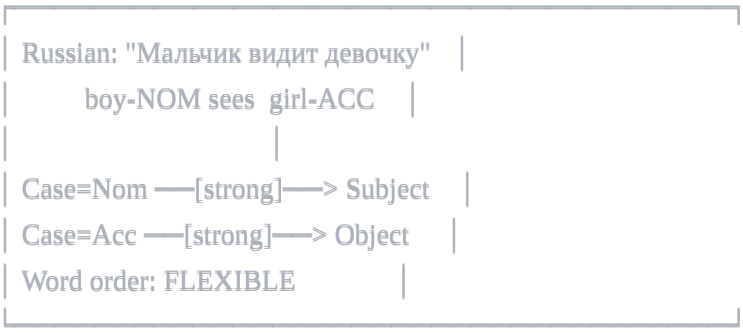
MORPHOLOGICAL FEATURES AS CHEMICAL PROPERTIES

Feature Compatibility Matrix

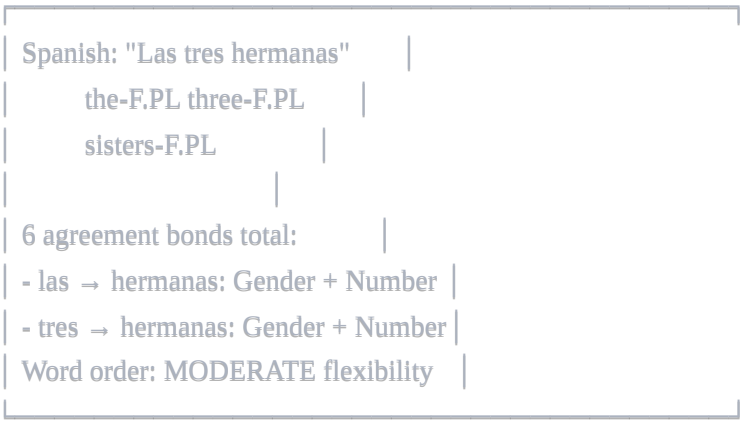
```
┌─────────────────────────────────────────────────────────────┐
│ STRONG ATTRACTION (form links readily)                      │
├─────────────────────────────────────────────────────────────┤
│ Case=Nom + Finite Verb      → Subject link (ionic bond)     │
│ Gender=Masc + Gender=Masc   → Agreement (H-bond)            │
│ Number=Sing + Number=Sing   → Agreement (H-bond)            │
│ Definite=Def + Prior mention → Anaphora (hydrophobic)       │
└─────────────────────────────────────────────────────────────┘

┌─────────────────────────────────────────────────────────────┐
│ NEUTRAL (no preference)                                     │
├─────────────────────────────────────────────────────────────┤
│ Case=Gen + Case=Nom         → No direct interaction         │
│ Gender=Masc + Gender=Fem    → No agreement required         │
└─────────────────────────────────────────────────────────────┘

┌─────────────────────────────────────────────────────────────┐
│ REPULSION (incompatible - prevent linking)                  │
├─────────────────────────────────────────────────────────────┤
│ Case=Nom + Object position  → BLOCKED                       │
│ Gender=Masc ↔ Gender=Fem    → Agreement violation           │
│ Number=Sing ↔ Number=Plur   → Agreement violation           │
│ VerbForm=Inf + Main clause  → Needs Fin for main clause     │
└─────────────────────────────────────────────────────────────┘
```

Cross-Linguistic Feature Profiles

CASE-BASED (Russian, Latin, Finnish):
Primary features: Case
Secondary: Gender, Number
Like: Strongly charged amino acids (Lys, Glu)
Effect: Case determines function regardless of position



AGREEMENT-BASED (Spanish, French, German):
Primary features: Gender, Number
Secondary: Case (reduced), Definiteness
Like: Polar amino acids with H-bonding
Effect: Multiple weak bonds create strong structure



POSITION-BASED (English, Chinese):
Primary features: Word order, Definiteness
Secondary: Limited agreement
Like: Hydrophobic effect (position matters)
Effect: Structure depends on position + information

COMPLETE EXAMPLE: End-to-End Processing

Input Sentence

Portuguese: "Tomei café da manhã cedo"
Translation: "I had breakfast early"

STAGE 1: TRANSCRIPTION

```
┌──────────────────────────────────────────────────────────┐
│ INPUT: Raw word sequence                                 │
│ "tomei café da manhã cedo"                               │
└──────────────────────────────────────────────────────────┘
                          ↓
┌──────────────────────────────────────────────────────────┐
│ UD FEATURE EXTRACTION                                    │
│ tomei:  VERB  VerbForm=Fin|Mood=Ind|                     │
│               Tense=Past|Person=1|Number=Sg              │
│ café:   NOUN  Gender=Masc|Number=Sing                    │
│ da:     ADP   (de+a contraction)                         │
│ manhã:  NOUN  Gender=Fem|Number=Sing                     │
│ cedo:   ADV   (no features)                              │
└──────────────────────────────────────────────────────────┘
                          ↓
┌──────────────────────────────────────────────────────────┐
│ TYPE CLASSIFICATION                                      │
│ tomei:  V (VerbForm=Fin → predicative)                   │
│ café:   E (NOUN → referential)                           │
│ da:     F (ADP → grammatical)                            │
│ manhã:  E (NOUN → referential)                           │
│ cedo:   A (ADV → attributive)                            │
└──────────────────────────────────────────────────────────┘
                          ↓
┌──────────────────────────────────────────────────────────┐
│ MWE DETECTION & ASSEMBLY                                 │
│ Check: "café da manhã" in MWE lexicon                    │
│ Found: Type=E, Components=[café,da,manhã]                │
│                                                          │
│ Activation sequence:                                     │
│ 1. café     : 1/3 (threshold=1) ✓ stable                 │
│ 2. café da  : 2/3 (threshold=2) ✓ stable                 │
│ 3. café da manhã: 3/3 (threshold=3) ✓ COMPLETE           │
│                                                          │
│ → AGGREGATE: Transfer links to MWE node                  │
│ → GARBAGE COLLECT: Remove sub-nodes                      │
└──────────────────────────────────────────────────────────┘
```

```
                ↓
┌──────────────────────────────────────────────────────┐
│ TRANSCRIPTION OUTPUT:                                │
│                                                      │
│ [1] tomei                                            │
│     Type: V                                          │
│     Features: VerbForm=Fin|Mood=Ind|...              │
│     Status: WORD_NODE                                │
│                                                      │
│ [2] café_da_manhã                                    │
│     Type: E                                          │
│     Features: Gender=Masc|Number=Sing                │
│     Status: MWE_NODE (aggregated)                    │
│                                                      │
│ [3] cedo                                             │
│     Type: A                                          │
│     Features: (none)                                 │
│     Status: WORD_NODE                                │
└──────────────────────────────────────────────────────┘

Croft annotation: "tomei + café^da^manhã + cedo"
```

STAGE 2: TRANSLATION

```
┌──────────────────────────────────────────────────────┐
│ INPUT: Stable lexical units                          │
│ [tomei:V] [café_da_manhã:E] [cedo:A]                 │
└──────────────────────────────────────────────────────┘
                ↓
┌──────────────────────────────────────────────────────┐
│ PHRASE BUILDING                                      │
│                                                      │
│ 1. tomei (V) → Start Predicate phrase                │
│    Predictions:                                      │
│    - Subject: type E/V (implicit in Portuguese)      │
│    - Object: type E                                  │
│                                                      │
│ 2. café_da_manhã (E) → Check predictions             │
│    Matches: Object prediction from tomei             │
│    Type check: E = E ✓                               │
│    Feature check: No conflicts ✓                     │
│    → CREATE LINK: tomei ──[OBJ]──> café_da_manhã     │
│                                                      │
│ 3. cedo (A) → Modifier                               │
│    Attaches to: tomei (temporal modifier)            │
│    → CREATE LINK: tomei ──[ADV]──> cedo              │
```

```
┌──────────────────────────────────────────────────────┐
└──────────────────────────────────────────────────────┘
                          ↓
┌──────────────────────────────────────────────────────┐
│ PHRASE LABELING (Croft's clausal CEs)                │
│                                                      │
│ [Pred: tomei]                                        │
│ [Arg:  café_da_manhã]                                │
│ [FPM:  cedo]                                         │
└──────────────────────────────────────────────────────┘
                          ↓
┌──────────────────────────────────────────────────────┐
│ TRANSLATION OUTPUT:                                  │
│                                                      │
│ Three separate phrases:                              │
│ • Predicate phrase: [tomei]                          │
│ • Argument phrase: [café_da_manhã]                   │
│ • Modifier phrase: [cedo]                            │
│                                                      │
│ Local links established:                             │
│ • tomei → café_da_manhã [OBJ]                        │
│ • tomei → cedo [ADV]                                 │
└──────────────────────────────────────────────────────┘
```

Croft annotation: "tomei + café^da^manhã + cedo ."
Labels: Pred + Arg + FPM

STAGE 3: FOLDING

```
┌──────────────────────────────────────────────────────┐
│ INPUT: Phrases with local links                      │
│ [Pred: tomei] [Arg: café_da_manhã] [FPM: cedo]       │
└──────────────────────────────────────────────────────┘
                          ↓
┌──────────────────────────────────────────────────────┐
│ SENTENCE INTEGRATION                                 │
│                                                      │
│ 1. Identify root: tomei (finite verb)                │
│                                                      │
│ 2. Build argument structure:                         │
│    - Subject: implicit (pro-drop)                    │
│    - Object: café_da_manhã (already linked)          │
│                                                      │
│ 3. Attach modifiers:                                 │
│    - cedo → tomei (temporal)                         │
│                                                      │
│ 4. No subordination (simple sentence)                │
```

```
│                                 │
│ 5. No long-distance dependencies │
│    (would create crossing edges in complex) │
└─────────────────────────────────────────────┘

                ↓

┌─────────────────────────────────────────────┐
│ FINAL PARSE GRAPH:                          │
│                                 │
│         tomei (ROOT)                        │
│         /    \                              │
│        /      \                             │
│  café_da_manhã  cedo                        │
│   [OBJ]        [ADV]                        │
│                                 │
│ Graph properties:                           │
│ - Nodes: 3 (all connected)                  │
│ - Edges: 2 (all labeled)                    │
│ - Projective: YES (no crossings)            │
│ - Complete: YES (all words linked)          │
└─────────────────────────────────────────────┘
```

Croft annotation: "tomei + café^da^manhã + cedo ."
Sentential: Main clause (simple sentence)

COMPLEX EXAMPLE: Long-Distance Dependencies

Input Sentence

Portuguese: "O menino que eu vi chegou cedo"
Translation: "The boy that I saw arrived early"

Key Challenge: Non-Projective Structure

Linear order: o_1 $menino_2$ que_3 eu_4 vi_5 $chegou_6$ $cedo_7$

Dependencies:
1. o → menino (determiner)
2. menino → chegou (subject) [CROSSES 3,4,5!]
3. que → menino (relative pronoun to antecedent)
4. eu → vi (subject of embedded clause)
5. que → vi (object of embedded clause)
6. cedo → chegou (adverb)

Graph structure (non-projective):

```
      chegou⁶ (ROOT)
      /   \
  menino²   cedo⁷
    |
  o¹ que³
      \
     vi⁵
     /
   eu⁴
```

The edge menino² → chegou⁶ crosses over que³, eu⁴, vi⁵
This is like a DISULFIDE BRIDGE in proteins!

Biological Parallel: Disulfide Bridge

Protein sequence:
...Cys_{15}-Ala_{16}-Gly_{17}-Val_{18}-Cys_{78}...

Linear: far apart
3D: close together via S-S bond

```
   Cys₁₅
    |
    S
    |
    S  ← Disulfide bridge
    |
   Cys₇₈
```

Relative clause:
$menino_2$ ... que_3 ... vi_5 ... $chegou_6$

Linear: separated by relative clause
Graph: direct link (crosses intervening words)

```
  menino₂
    |
  (subject)
    |
  chegou₆  ← Direct dependency
```

Croft's Representation

"O menino {que + eu + vi} + chegou + cedo ."

Boundary interpretation:
- Main clause: [O menino] [chegou] [cedo]
- Embedded (relative): {que eu vi}
- Interruption: {} marks center-embedding

Phrasal CEs:
o: Mod (of menino)
menino: Head
que: Head (of relative)
eu: Head
vi: Head
chegou: Head
cedo: Head

Clausal CEs:
o menino: Arg (of chegou)
que eu vi: Rel (modifies menino)
chegou: Pred (main)
cedo: FPM (modifier)

Sentential CEs:
Main clause: [o menino chegou cedo]
Relative clause: {que eu vi}

PRACTICAL IMPLEMENTATION CHECKLIST

Phase 1: Core Architecture ✓

- [ ] Create TranscriptionService.php
- [ ] Create TranslationService.php
- [ ] Create FoldingService.php
- [ ] Refactor ParserService.php to orchestrate stages
- [ ] Add stage logging

Phase 2: Feature Extraction ✓

- [ ] Extend UDParserService for full features
- [ ] Add feature storage (JSONB columns)
- [ ] Create FeatureBundle data class
- [ ] Test feature extraction accuracy

Phase 3: Feature-Driven Linking ✓

- [ ] Implement FeatureCompatibilityService
- [ ] Add agreement checking
- [ ] Add case compatibility
- [ ] Add definiteness handling
- [ ] Test with multiple languages

Phase 4: Advanced Operations ✓

- [ ] Implement long-distance dependencies
- [ ] Handle non-projective structures
- [ ] Add feature propagation
- [ ] Update visualizations

Phase 5: Evaluation ✓

- [ ] Create test corpus (multiple languages)
- [ ] Implement evaluation metrics
- [ ] Compare against baseline
- [ ] Document results

QUICK REFERENCE: Feature → Chemistry Mappings

UD Feature	Chemical Property	Effect
Case=Nom	Positive charge	Attracts Subject
Case=Acc	Negative charge	Attracts Object
Gender=Masc/Fem	Polarity	Agreement
Number=Sing/Plur	Size	Agreement
Definite=Def	Hydrophobic	Clusters
VerbForm=Fin	Catalytic site	Predicative
Mood=Ind	pH neutral	Assertion
Tense=Past	Time marker	Reference
Person=1/2/3	Identity	Agreement

Document Type: Visual Reference Guide
Version: 1.0

Date: December 2024
Companion to: transdisciplinary_parsing_research.md