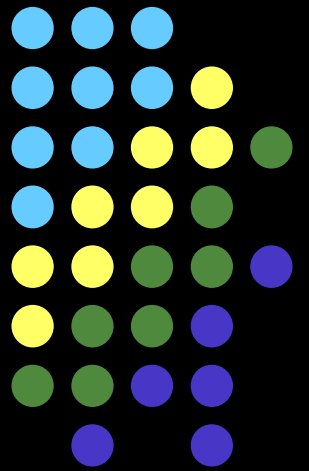
# Projetando o *corpus* para a construção de uma *wordnet* terminológica

**Ariani Di Felippo**

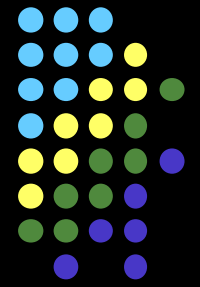
**Jackson W. da Cruz Souza**

Departamento de Letras - DL

Universidade Federal de São Carlos - UFSCar

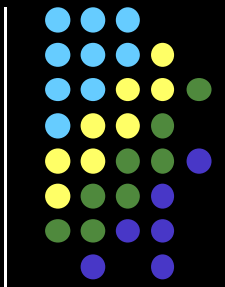


# Contexto



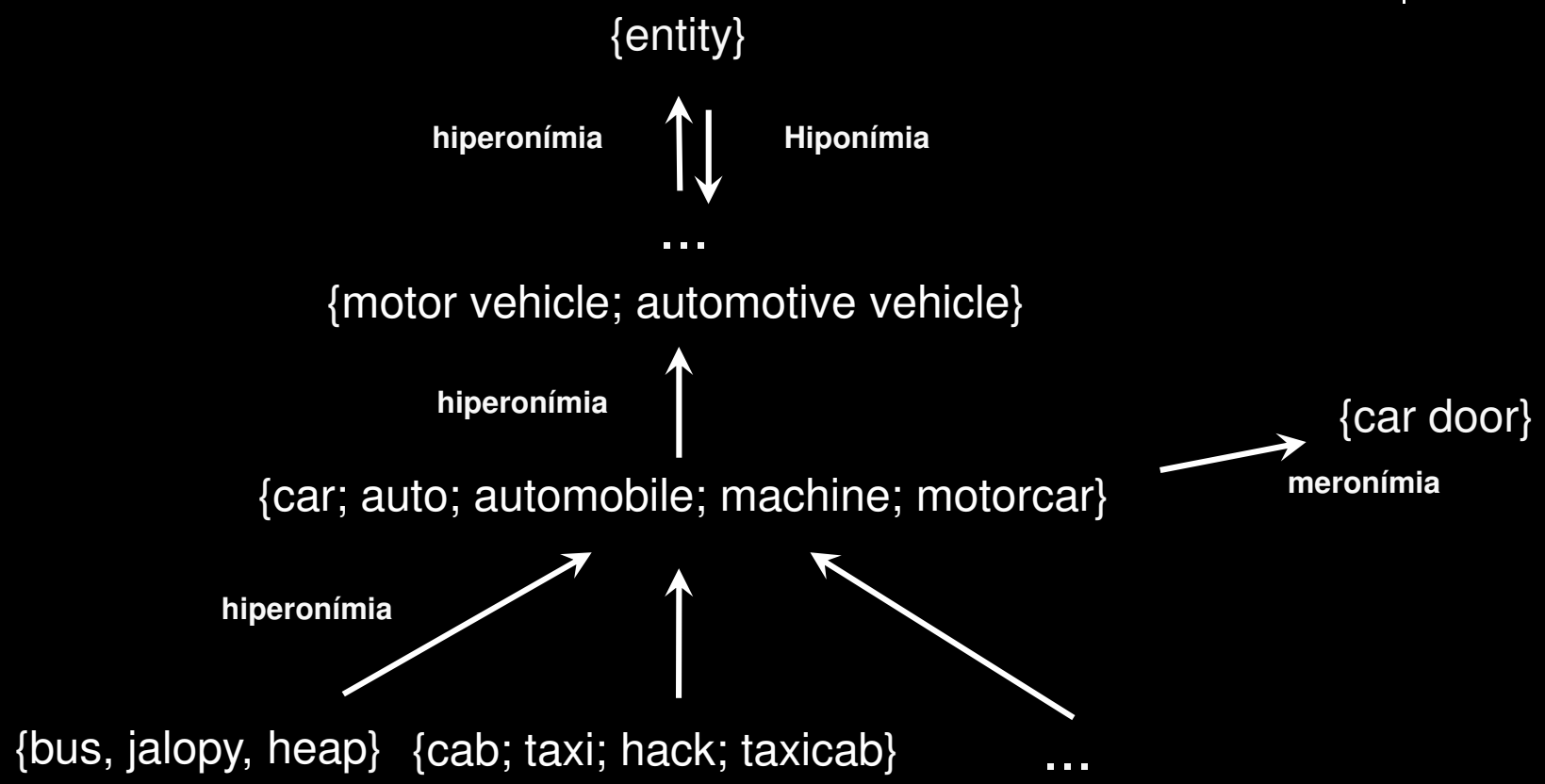
- Sistemas computacionais que processam língua natural
  - Baseados em conhecimento linguístico
    - Bases de dados lexicais (BDLs)
- Formato *wordnet* para BDLs
  - Princeton WordNet (Fellbaum, 1998)

# As *wordnets*

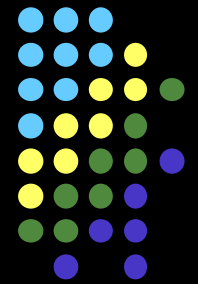


- **Estrutura**
  - **Categoria sintática: V, N, Adj e Adv**
  - **Conjunto de sinônimos: "synonym set" → *synset***
    - **{dog, domestic dog, Canis familiaris}**
  - **Relações conceituais** (entre *synsets*)
    - **hiponímia/ hiperonímia, meronímia/ holonímia, acarretamento e causa**

# As *wordnets*

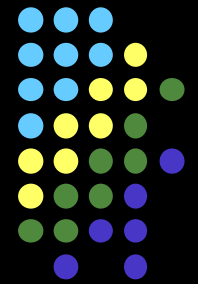

{entity}
hiperonímia
Hiponímia
...
{motor vehicle; automotive vehicle}
hiperonímia
{car door}
meronímia
{car; auto; automobile; machine; motorcar}
hiperonímia
{bus, jalopy, heap}
{cab; taxi; hack; taxicab}
...

# Contexto



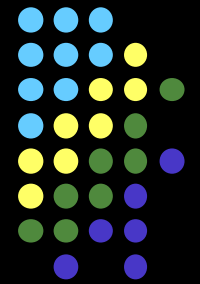
- **Processamento de textos técnicos**
  - ***Wordnets* terminológicas**
    - **JurWordnet** (Sagri et al., 2004)
    - **ArchiWordnet** (Bentivogli et al., 2004)
    - **Medical Wordnet** (Smith, Fellbaum, 2004)
    - **BioWordnet** (Poprat et al., 2008)
- **Ausência de uma metodologia clara e genérica**

# Contexto



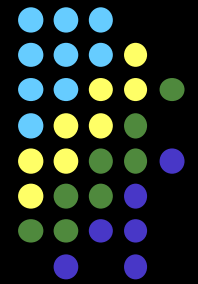
- Projeto **TermiNet**
  - 1º objetivo
    - **Instanciar** a metodologia genérica de pesquisa no PLN (Dias-da-Silva, 2006) para o desenvolvimento de *wordnets* terminológicas (ou *terminets*) em PB
      - Domínio linguístico
      - Domínio representacional
      - Domínio implementacional

# Contexto



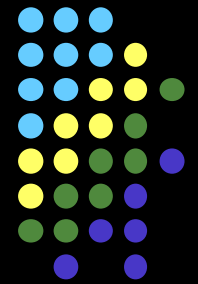
- Projeto **TermiNet**
  - 2º objetivo
    - **Validar** a metodologia instanciada por meio da construção de uma terminet em PB (protótipo)
      - Educação à Distância

# Contexto



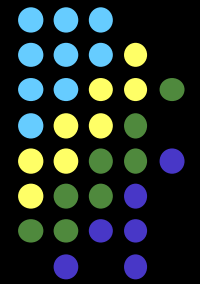
- **Ao instanciar a metodologia genérica ...**
  - **Domínio linguístico**
    a) **Delimitação do domínio especializado**
    b) **Delimitação das fontes para a aquisição do conhecimento léxico-conceitual característico de uma *wordnet***
    c) **Compilação do conhecimento léxico-conceitual**

# Contexto



- **Delimitação das fontes**
  - **Estruturadas**
    - **Dicionários, *thesauri,* taxonomias, etc.**
  - **Não-estruturadas**
    - ***Corpora***
      - *"A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research."* (Sinclair, 2005)

# Contexto



- As **etapas** de construção de um *corpus* (Almeida, Aluísio, 2007):

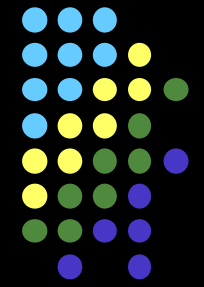
(a) **Projeto** do *corpus* → definição do tipo de *corpus* necessário à pesquisa

(b) **Compilação**

(c) **Pré-processamento** → conversão, limpeza, nomeação e anotação

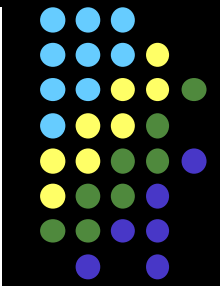
(d) **Aquisição das permissões de uso** (caso seja disponibilizado na *Web*).

# Premissa



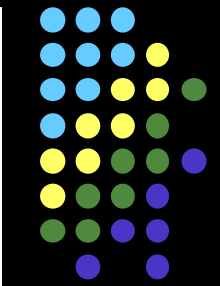
- O projeto do *corpus* **depende** de 3 fatores
  - **Requisitos** ou critérios que definem "*corpus*"
  - **Recurso lexical** que será construído a partir do *corpus*
  - **Decisões de projeto**

# O projeto do *corpus*



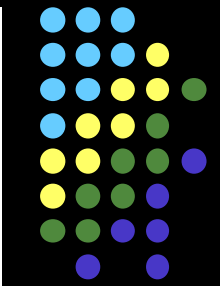
- **Requisitos** (cf. Kennedy (1998), Biber et al. (1998), Renouf (1998), Sardinha (2004) e Sinclair (2005))
  - **Representatividade/ Amostragem**
    - **Um *corpus* deve ter uma amostragem suficiente da língua ou variedade de língua que se quer analisar para se obter o máximo de representatividade dessa mesma língua ou variedade de língua**

# O projeto do *corpus*



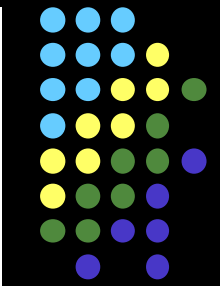
- **Requisitos** (cf. Kennedy (1998), Biber et al. (1998), Renouf (1998), Sardinha (2004) e Sinclair (2005))
  - **Tamanho**
    - **Todo *corpus* ter um tamanho finito** (com exceção de um *corpus* monitor)

# O projeto do *corpus*



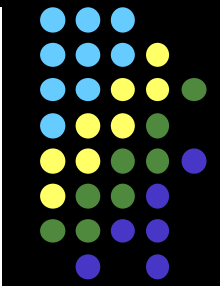
- **Requisitos** (cf. Kennedy (1998), Biber et al. (1998), Renouf (1998), Sardinha (2004) e Sinclair (2005))
  - **Autenticidade**
    - **Um *corpus* deve conter textos que existem na linguagem, ou seja, que não foram criados com o objetivo de figurarem no *corpus***

# O projeto do *corpus*



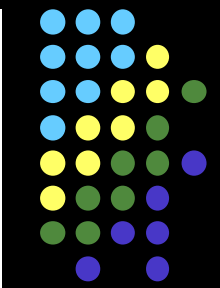
- **Requisitos** (cf. Kennedy (1998), Biber et al. (1998), Renouf (1998), Sardinha (2004) e Sinclair (2005))
  - **Diversidade/ Balanceamento**
    - **A quantidade de textos deve estar equilibrada em função dos gêneros discursivos, tipos de textos, etc., desde que as escolhas sejam adequadas à pesquisa que se pretende realizar**

# O projeto do *corpus*



- 1ª pergunta:
  - Como satisfazer os requisitos que formam a essência de um *corpus*?

# O projeto do *corpus*

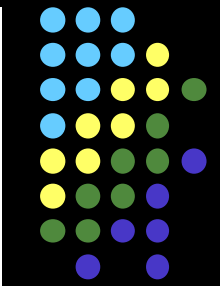


- **Representatividade/ Amostragem / Tamanho**
  - **Construção de um *corpus* médio-grande**

    (+ 1 milhão de palavras)
    - Nascimento (2003), Aluísio e Almeida (2007), Coleti et al. (2008)

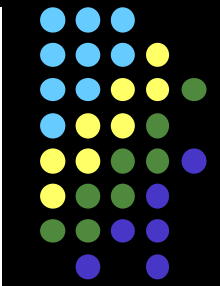
- **Autenticidade**
  - **Coleta de textos em comunicações "especializadas" genuínas e de fontes confiáveis**
    - **de preferência, textos escritos por falantes nativos**

# O projeto do *corpus*



- **Diversidade**
  - **Coleta de textos dos gêneros técnico-científico, científico de divulgação, instrucional, informativo e técnico-administrativo**
  - **Textos veiculados por livros, revistas, jornais, manuais, etc.**
    - Nascimento (2003) e Agbago e Barrière (2005)
- **Balanceamento**
  - **Distribuição equilibrada dos gêneros textuais**

# O projeto do *corpus*

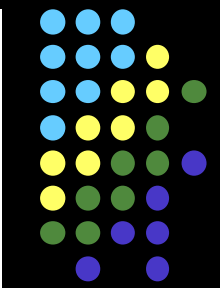


- 2ª pergunta:
  - Quais seriam as **características** diretamente dependentes do recurso a ser construído, ou seja, de uma *terminet* em PB?

# O projeto do *corpus*

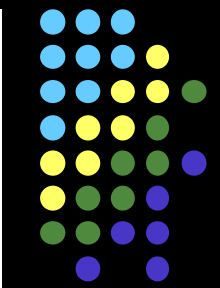
- Tendo em vista a construção de uma *terminet*, o *corpus* deve conter…
  - Textos em PB → monolíngue
  - Textos relativos ao domínio para o qual a *terminet* está sendo construída → especializado
  - Textos que contenham linguagem contemporânea, proporcionando a descrição sincrônica do léxico do domínio em questão → sincrônico

# O projeto do *corpus*



- Tendo em vista a construção de uma *terminet*, o *corpus* deve conter…
  - Textos registrados em meio escrito (digitais ou impressos), pois as *terminets* são recursos para o tratamento computacional das línguas naturais registradas em tal meio → escrito (meio)
  - * Textos de língua escrita, devido à dificuldade de aquisição de material transcrito → escrito (modalidade)

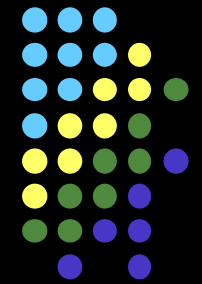
# O projeto do *corpus*



- 3ª pergunta:
  - Quais seriam as **características** diretamente dependentes de decisões de projeto?

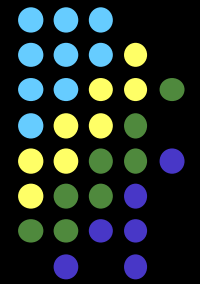
- Tendo em vista a aplicação de alguns métodos semiautomáticos de extração de conhecimento, o *corpus* deve ser...
  - anotado morfossintaticamente

# O projeto do *corpus*



- Tendo em vista que um *corpus* especializado é um recurso útil e de construção cara, ele deve ser…
  - disponível (via *Web*)
- Tendo em vista a construção de um recurso específico (*terminet*), o *corpus,* uma vez construído, não será modificado e, por tanto...
  - fechado

# Tipologia



| Tipologia de Gioulli e Peperidis (2007) | |
|---|---|
| **Modalidade** | Escrito |
| **Tipo de texto** | Escrito (língua escrita registrada em meio escrito) |
| **Mídia** | Jornais, livros, manuais, periódicos e outras |
| **Cobertura da língua** | Especializado |
| **Gênero** | Técnico-científico, científico de divulgação, instrucional, informativo e técnico-administrativo |
| **Quantidade de línguas** | Monolíngue |
| **Anotação** | Anotado (nível morfossintático) |
| **Comunidade produtora** | Falantes nativos |
| **Mutabilidade** | Fechado |
| **Variação histórica** | Sincrônico (contemporâneo) |
| **Disponibilidade** | Disponível via *Web* |

# OBRIGADA!