UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Desenvolvimento de um recurso léxico com papéis semânticos para o português

Leonardo Zilio (PPG-Letras/UFRGS)

Orientadora: Profª. Drª. Maria José Bocorny Finatto (PPG-Letras/UFRGS)

Coorientadora: Profª. Drª. Aline Villavicencio (PPGC/UFRGS)

Orientador no Exterior: Prof. Dr. Mathieu Mangeot-Nagata (LIG/UJF)

Coorientador no Exterior: Prof. Dr. Carlos Ramisch (LIF/AMU)

CAMELEON

CAPES

CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico
60 ANOS

UFRGS
PROGRAMA DE PÓS-GRADUAÇÃO EM LETRAS

Informações Gerais

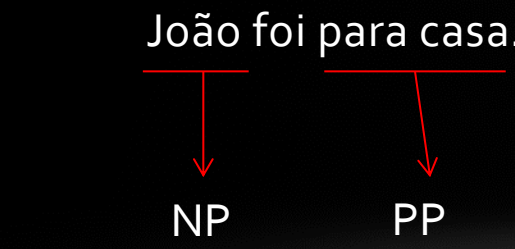
Uso de corpora especializado e não-especializado

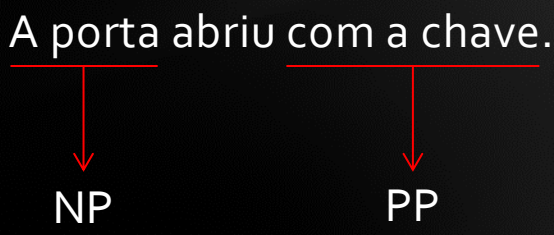
Trabalho realizado no nível da oração

Anotação manual de informações semânticas

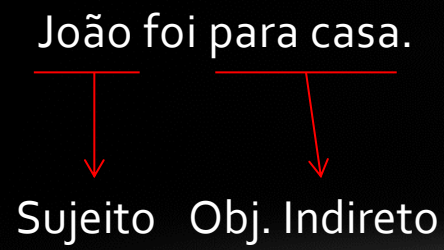
Essa anotação é feita com base em dois tipos de anotação já existentes

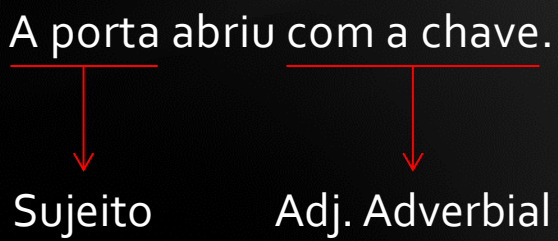
Anotação de Subcategorização



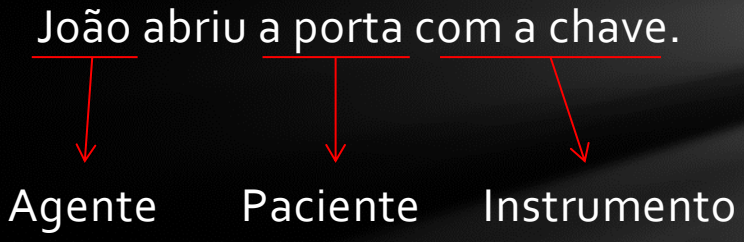


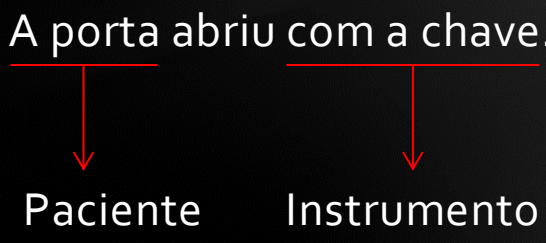
Anotação Sintática





Anotação de Papéis Semânticos





Objetivos

Gerar um recurso léxico com anotação de papéis semânticos

Contrastar as estruturas argumentais de verbos de linguagem especializada com os de linguagem não especializada

Exemplo Especializado vs. Não Especializado

Corpus de Cardiologia:

- *"Atualmente esse aparelho pode ser encontrado nas unidades de atendimento, porém sua interpretação depende de especialistas, que muitas vezes não se encontram presentes em o momento de o exame."*
  - Estrutura: NP/tema + VP + PP/local

Corpus DG:

- *"O pé direito do calçado foi encontrado no buraco da loja de celulares."*
  - Estrutura: NP/tema + VP + PP/local

Os elementos linguísticos presentes são diferentes, mas a estrutura de papéis semânticos é a mesma.

Corpora

Corpus de Cardiologia

Compilado por Zilio (2009)

490 artigos especializados extraídos de 3 periódicos brasileiros

> 1,5 milhão de palavras

Anotado com o *parser* PALAVRAS (Bick, 2000)

Corpus Diário Gaúcho

Compilado no âmbito do projeto PorPopular

(http://www6.ufrgs.br/textecc/porlexbras/porpopular/index.php)

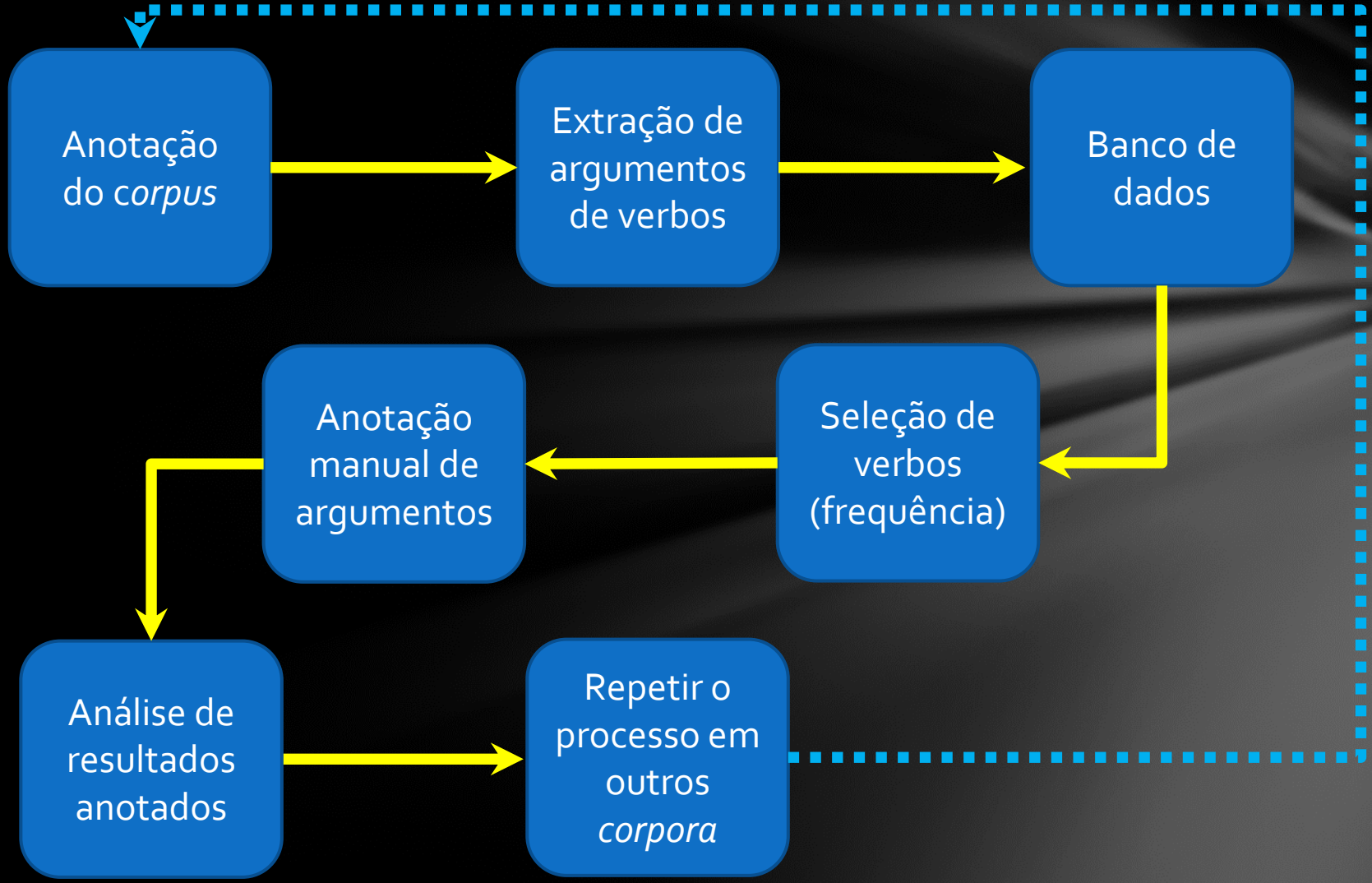
Textos retirados do jornal Diário Gaúcho (2008)

> 1 milhão de palavras

Anotado com o *parser* PALAVRAS (Bick, 2000)

Metodologia

Etapas do Trabalho



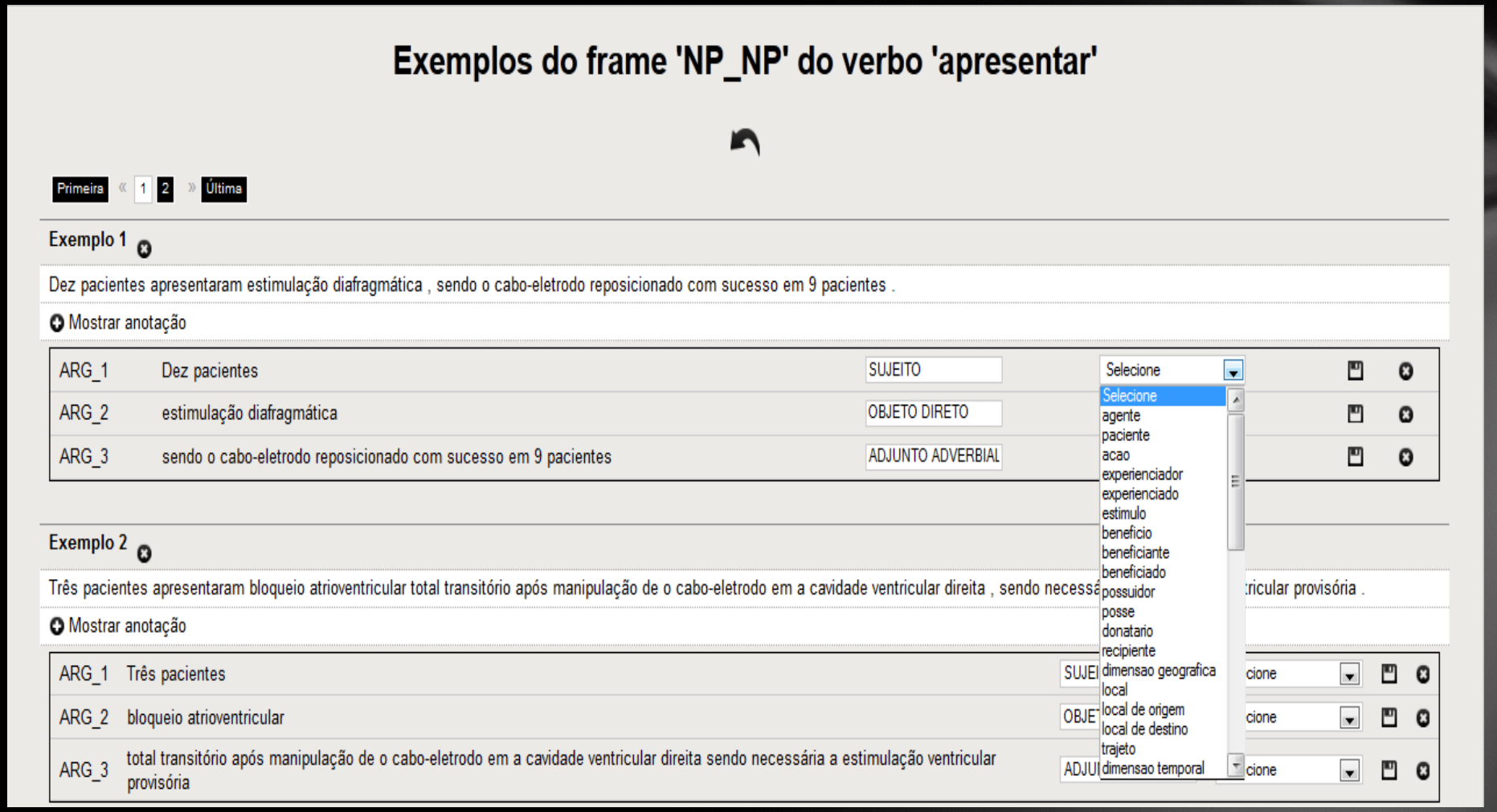
Extração automática de estruturas argumentais

Zanette (2010) - Extrator de *subcategorization frames* verbais a partir das árvores deitadas do PALAVRAS (BICK, 2000)

2011 = Modificação para árvore de dependências

Zanette et al. (2012) e Zilio et al. (2012) = Modificações nos parâmetros de reconhecimento de argumentos

A anotação de papéis semânticos



Doutorado-Sanduíche

Estudo-Piloto

50 verbos manualmente anotados nos dois *corpora*

Exportação dos dados para XML

Trabalho de Master I: Aluno Samy Sassi

Conversão de MySQL para XML

Comparação das anotações nos dois *corpora*

Dados da Anotação

3.400 Sentenças anotadas

- 1.790 de Cardiologia
- 1.610 do Diário Gaúcho

Cardiologia = 304 estruturas

Diário Gaúcho = 272 estruturas

Estruturas mais frequentes nos dois *corpora*

Diário Gaúcho		
Estrutura	Freq.	Freq. %
SUJ<Agent>+OBJ.DIR<Theme>	171	10,62
SUJ<Theme>	114	7,08
SUJ<Agent>	92	5,71
SUJ<Theme>+ADJ.ADV [em] <Location>	50	3,11
SUJ<Agent>+OBJ.DIR<Theme> +ADJ.ADV [em]<Location>	45	2,79

Cardiologia		
Estrutura	Freq.	Freq. %
SUJ<Theme>	181	10,11
SUJ<Theme>+ADJ.ADV[em] <Location>	121	6,76
SUJ<Instrument>+OBJ.DIR <Theme>	102	5,70
SUJ<Agent>+OBJ.DIR<Theme>	63	3,52
SUJ<Patient>	40	2,23

Comparação Quantitativa

Coeficiente Tau-b de Kendall:

Correlação inexistente para estruturas sintático-semânticas ($\tau = 0{,}031$).

SUBJ<Agent>+DIR.OBJ<Theme>

SUBJ<Experiencer>+DIR.OBJ <Theme>

Correlação existente para argumentos ($\tau = 0{,}523$):

SUBJ<Agent>

SUBJ<Experiencer>

DIR.OBJ<Theme>

Estudo-Piloto: Comparação Qualitativa

Maior uso de voz passiva e de intransitivos no *corpus* de Cardiologia

Maior presença de Instrumentos na posição de sujeito no *corpus* de Cardiologia

Pra que serviu?

Aprimorar a lista de papéis semânticos para dar conta de mais tipos de argumentos

Ter uma ideia preliminar da diferença/semelhança entre as linguagens comum e especializada do ponto de vista dos papéis semânticos

Atividades paralelas

Teste com vários anotadores

Transferência dos dados para a plataforma Jibiki

http://jibiki.univ-savoie.fr/jibiki/Home.po

Teste com vários anotadores

10 anotadores linguistas com treinamento básico

25 sentenças extraídas do *corpus*

Cálculo de concordância entre anotadores (multi-π)

Baixa concordância:

$\pi = 0{,}253407$

Plataforma Jibiki

http://jibiki.univ-savoie.fr/jibiki/Home.po

Plataforma Jibiki

Plataforma Jibiki

Estrutura: SUBJ_V_NP Voz: ATIVA Freq: 70

- Encontrei um túmulo destruído , que não tinha dono , com os dois vasos .

  Argumentos

  - *OCULTO*
    Sintaxe: SUJEITO Papel Semântico: agente Papel VerbNet:
  - *um túmulo destruído que não tinha dono com os dois vasos*
    Sintaxe: OBJETO DIRETO Papel Semântico: tema Papel VerbNet:

- Anderson beija Fernanda , quando encontra Klaus e Luiza .

  Argumentos

  - *OCULTO*
    Sintaxe: SUJEITO Papel Semântico: Papel VerbNet:
  - *Klaus Luiza*
    Sintaxe: OBJETO DIRETO Papel Semântico: Papel VerbNet:

- Isso aqui é a melhor terapia que eu poderia ter encontrado .

  Argumentos

  - *que*
    Sintaxe: OBJETO DIRETO Papel Semântico: Papel VerbNet:
  - *eu*
    Sintaxe: SUJEITO Papel Semântico: Papel VerbNet:

- Encontrei um bom acerto de o carro , gosto de essa pista e acho que seremos competitivos afirmou Felipe .

Andamento

Modificação na lista de Papéis Semânticos

86 verbos anotados até agora

Modificação da interface da plataforma Jibiki

Qualificação do Doutorado

Bibliografia

- BICK, Eckhardt. (2000) *The Parsing System* PALAVRAS*: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus: Aarhus University Press. Disponível em: http://beta.visl.sdu.dk/~eckhard/pdf/PLP20-amilo.ps.pdf
- ZANETTE, Adriano. (2010) *Aquisição de* Subcategorization Frames *para Verbos da Língua Portuguesa*. Projeto de Diplomação. UFRGS. Orientadora: Aline Villavicencio.
- ZANETTE, A; SCARTON, C.; ZILIO, L. (2012) Automatic extraction of subcategorization frames from corpora: an approach to Portuguese. In: *Proceedings of PROPOR 2012* - Demonstration Session. Coimbra, Portugal.
- ZILIO, Leonardo. (2009) *Colocações especializadas e* Komposita*: um estudo contrastivo alemão-português na área de Cardiologia*. Dissertação de Mestrado. Orientadora: Maria José Bocorny Finatto. Disponível em: http://www.lume.ufrgs.br/bitstream/handle/10183/16877/000706196.pdf?sequence=1
- ZILIO, Leonardo; ZANETTE, Adriano; SCARTON, Carolina (2012) Automatic extraction of subcategorization frames from corpora in portuguese. In: XI Encontro de Linguística de Corpus, São Carlos, SP, Brasil, 11-15 de setembro de 2012.

OBRIGADO!

Leonardo Zilio
ziliotradutor@gmail.com

Anotação do PALAVRAS

João viu o cachorro.

```
João        [João] <hum> PROP M S @SUBJ>  #1->2
viu         [ver] <vH> <fmc> <mv> V PS 3S IND VFIN @FS-STA  #2->0
o           [o] <artd> DET M S @>N  #3->4
cachorro    [cachorro] <Azo> N M S @<ACC  #4->2
$.                     #5->0
</s>
```

- Forma canônica
- Informação sintática
- Informação complementar
- Anotação da dependência
- Informação gramatical

Papéis Semânticos/Temáticos/Theta ou Case Frames

Fillmore (1968) – Case Frame

Jackendoff (1987, apud Reinhart, 2002) – Relações Temáticas

Von Polenz (1988, apud Gelhausen, 2010) – Papéis Temáticos

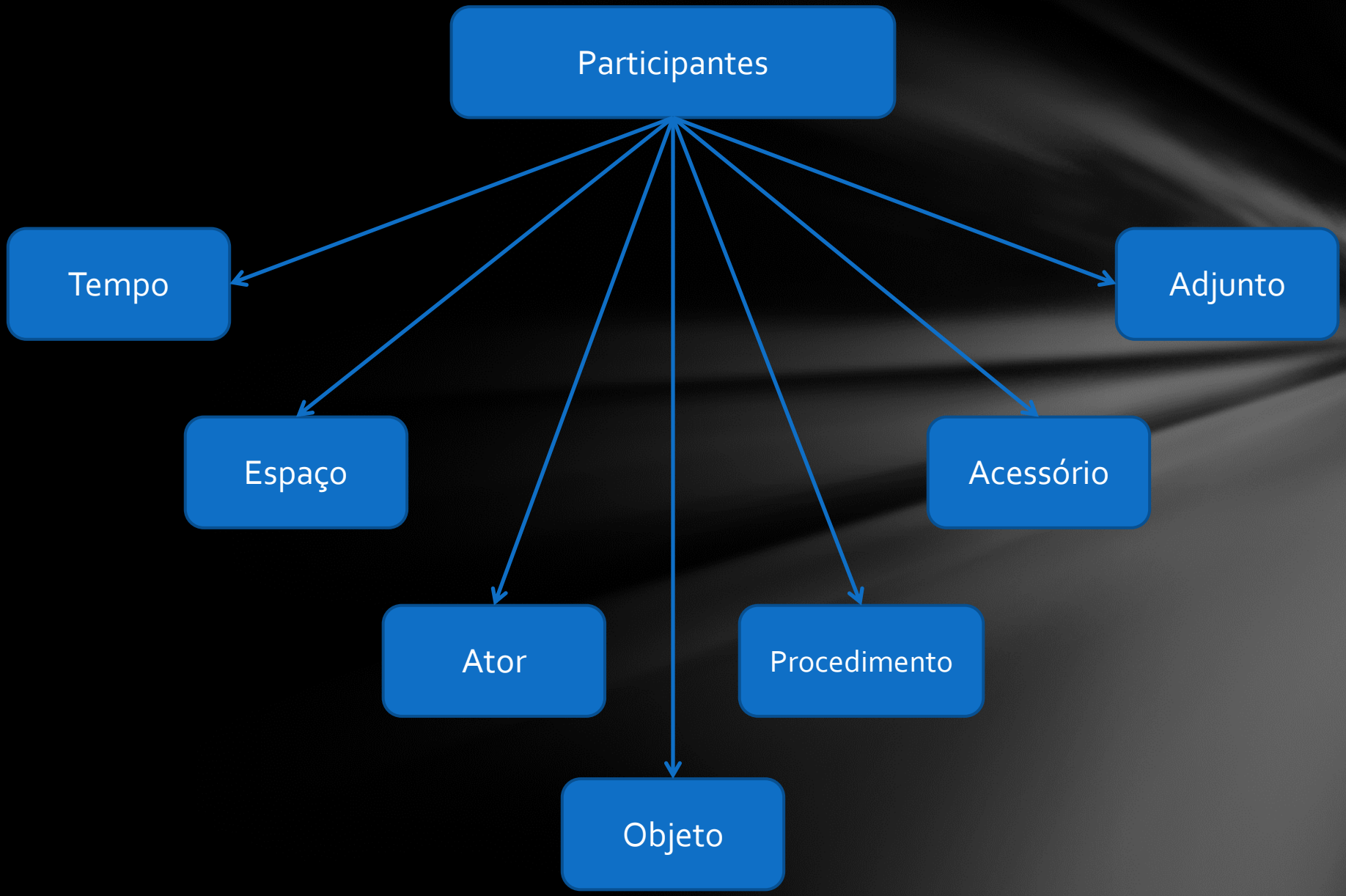
Levin & Rappaport-Hovav (2005) – Papéis Semânticos

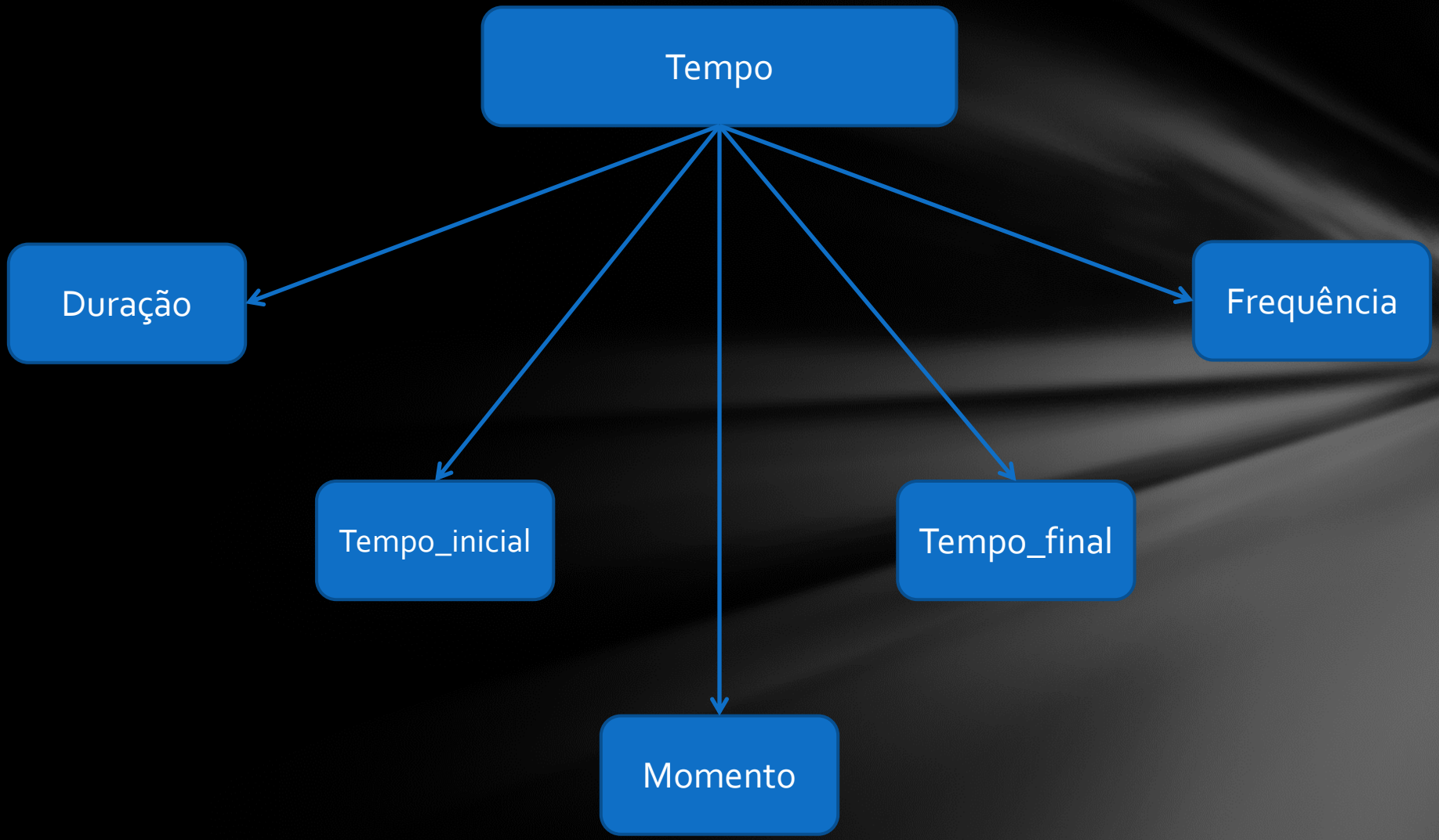
Gramática Gerativa – Theta Roles (Reinhart, 2002)

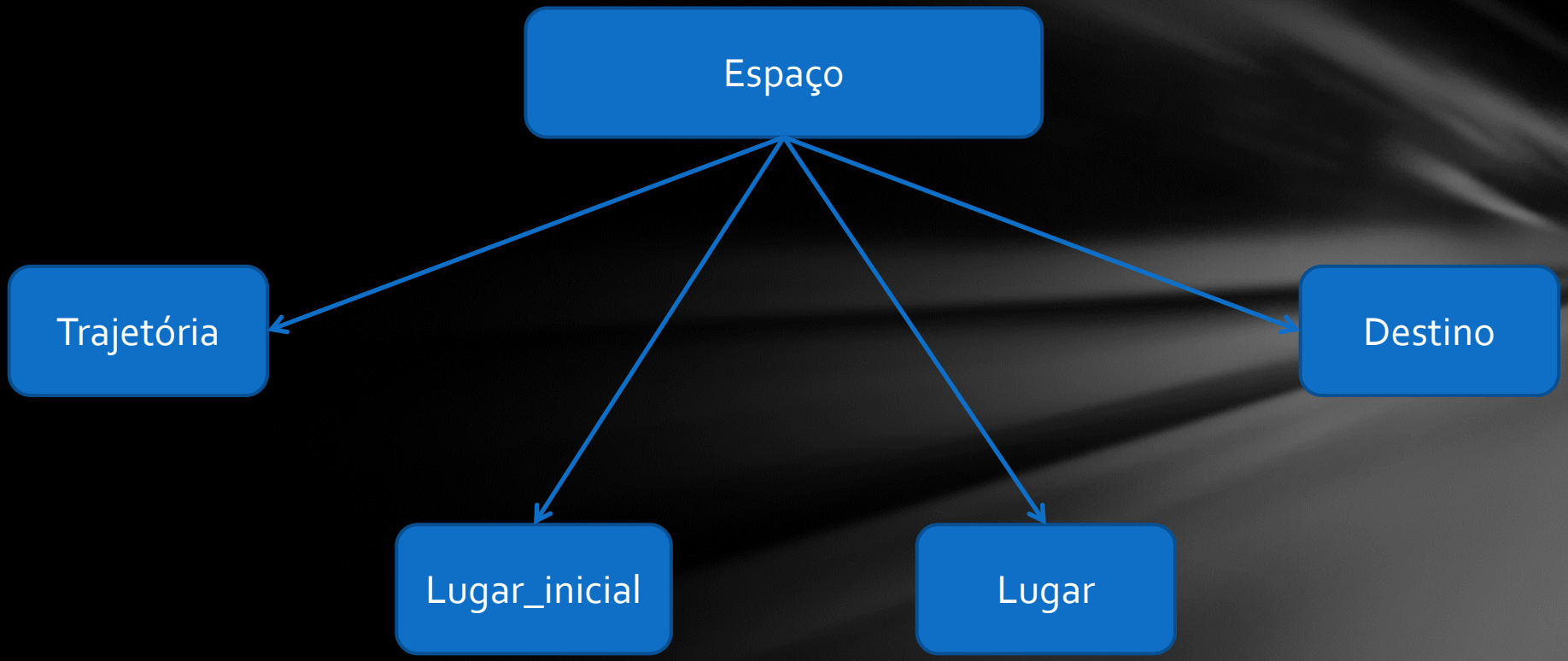
Dowty (1991) – Protopapéis Temáticos

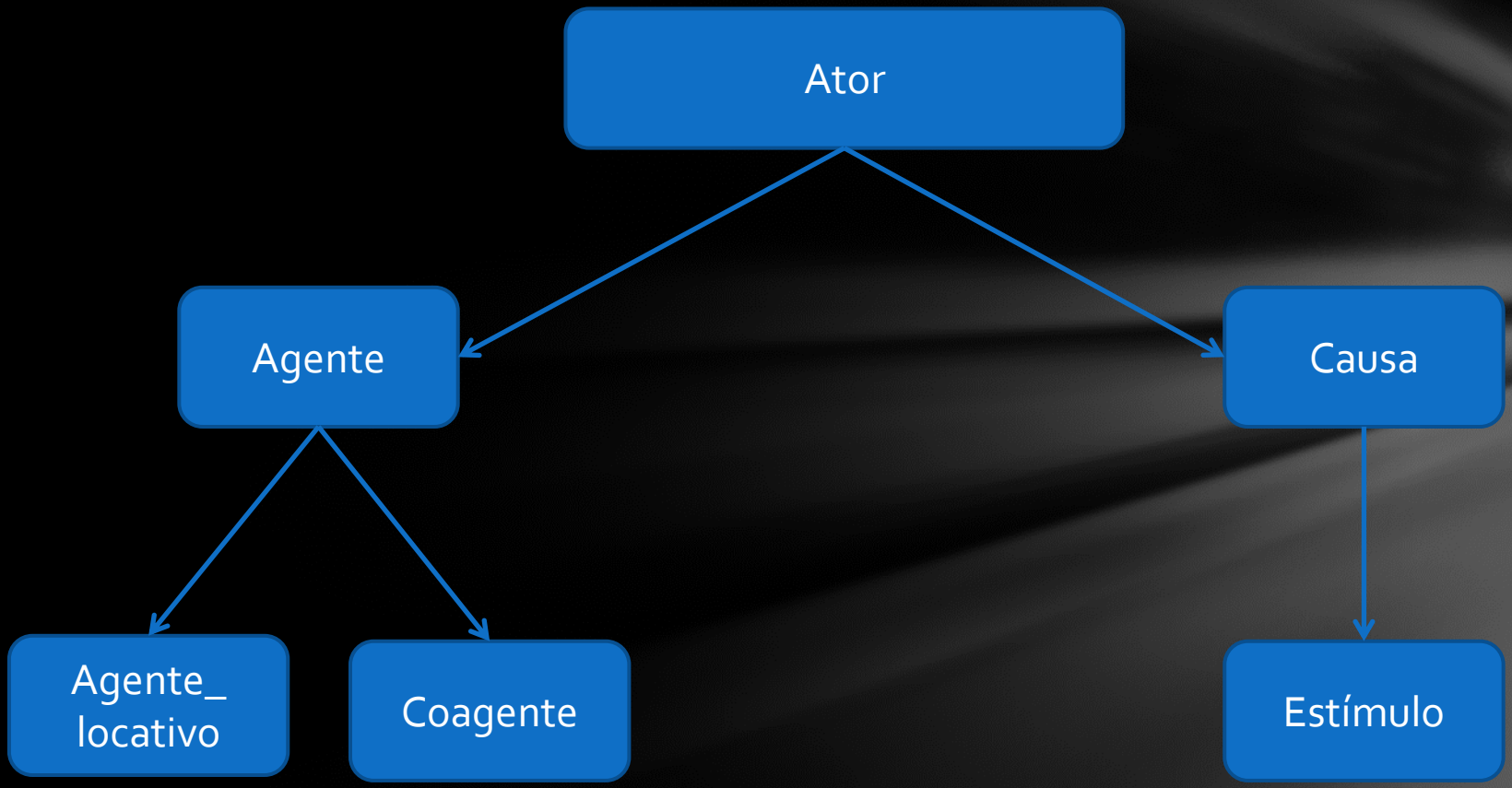
Papéis Semânticos

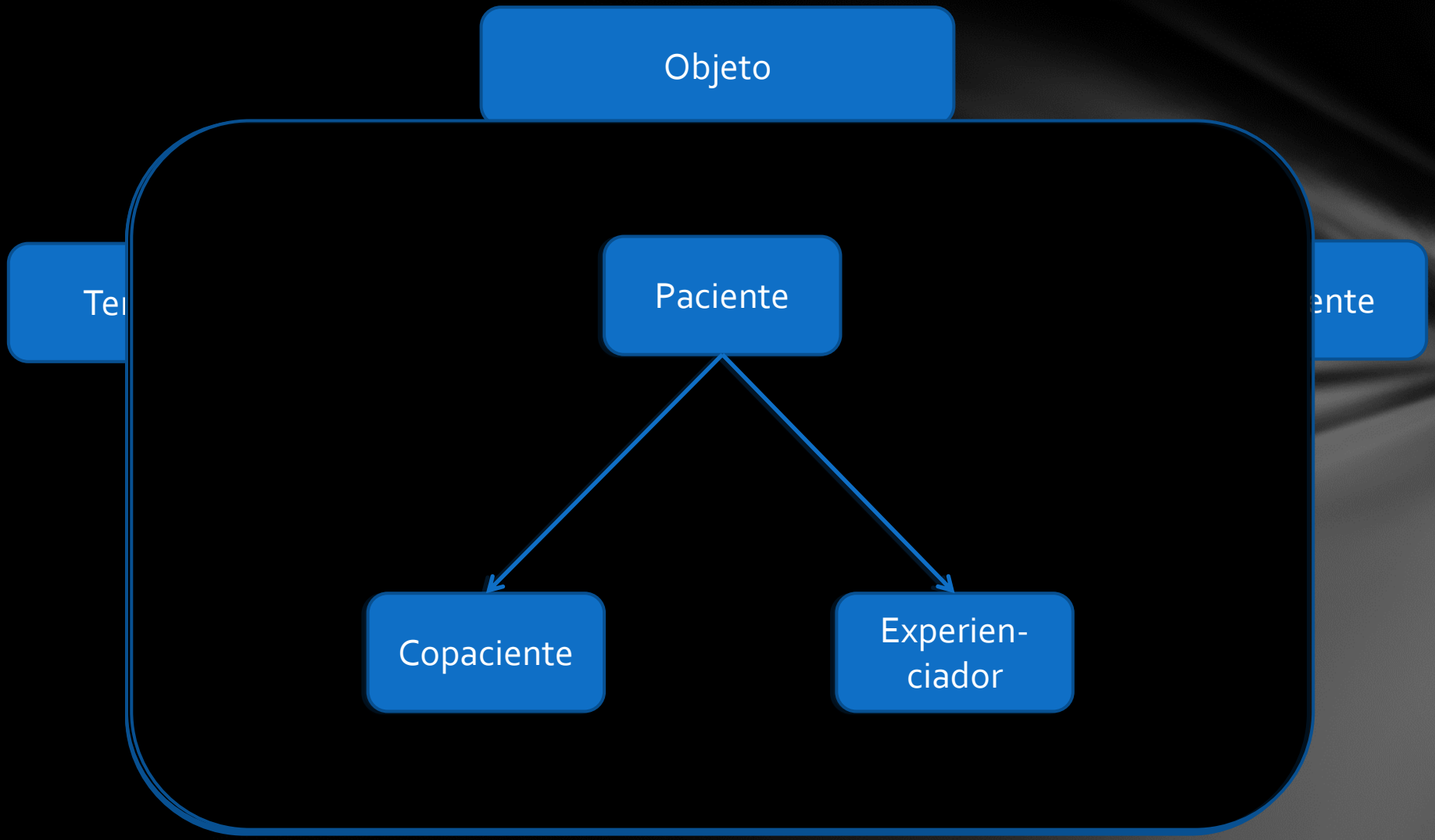
Lista final estendida

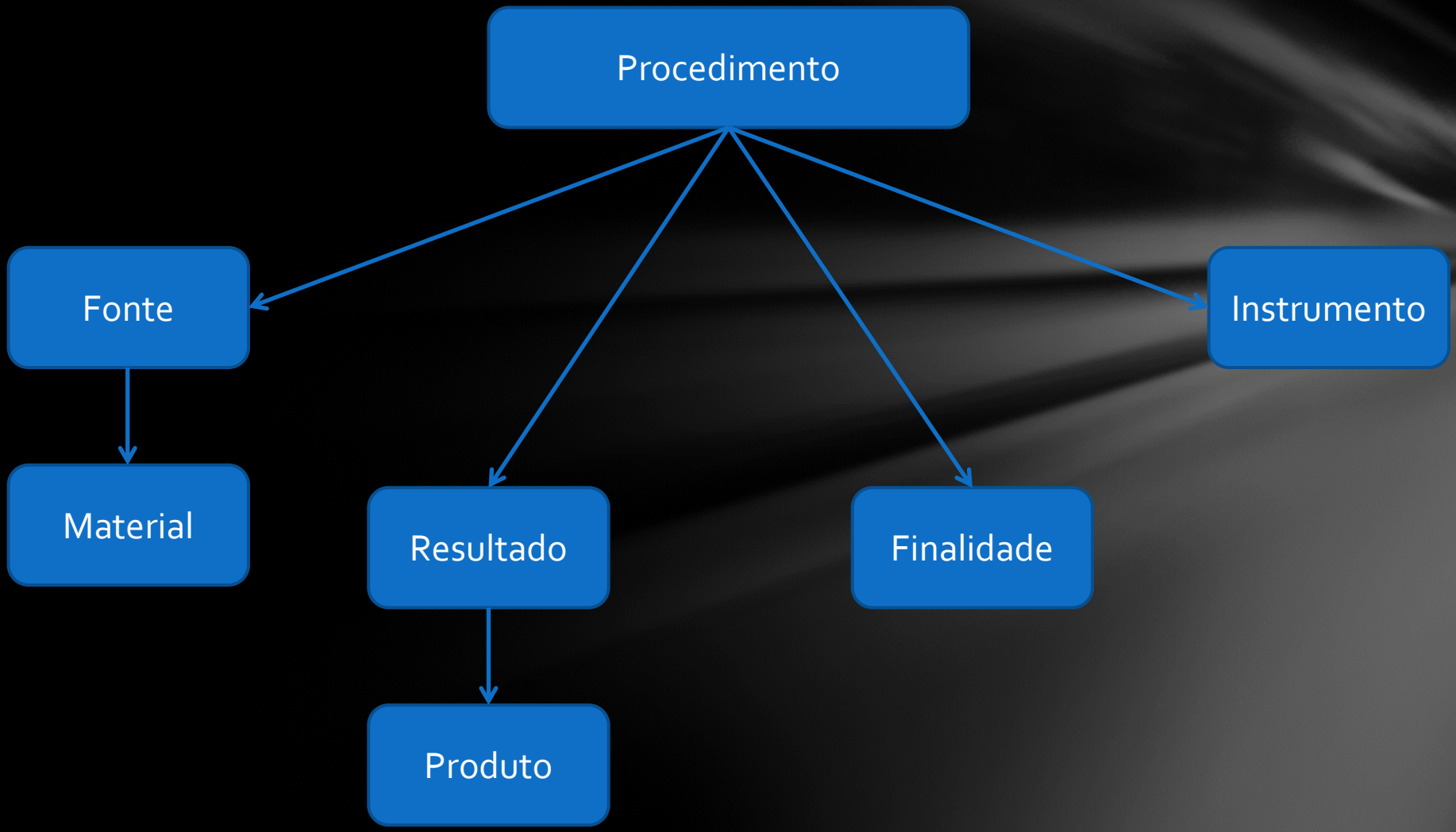

Participantes
Tempo
Adjunto
Espaço
Acessório
Ator
Procedimento
Objeto

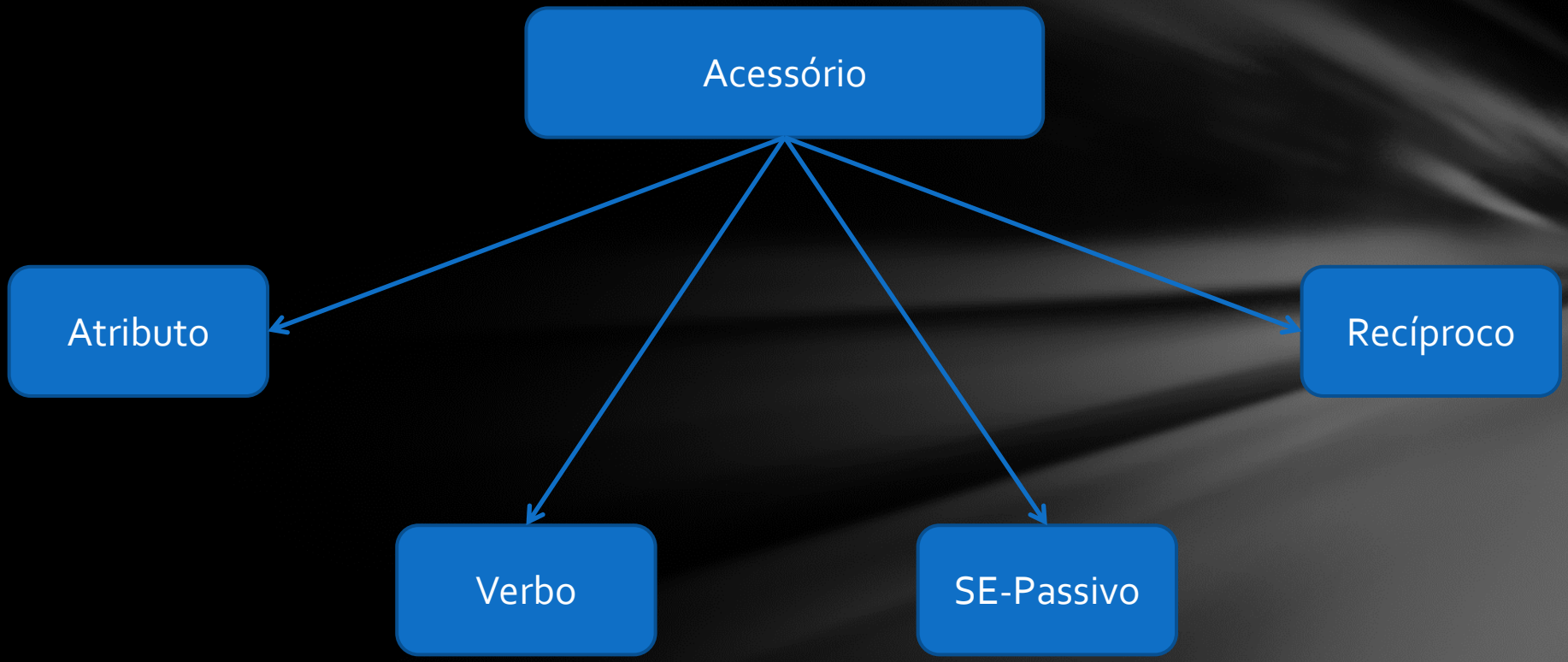

Tempo
Duração
Frequência
Tempo_inicial
Tempo_final
Momento


Espaço
Trajetória
Destino
Lugar_inicial
Lugar


Ator
Agente
Causa
Agente_
locativo
Coagente
Estímulo


Objeto
Ter
Paciente
ente
Copaciente
Experien-
ciador


Procedimento
Fonte
Material
Resultado
Produto
Finalidade
Instrumento


Acessório
Atributo
Recíproco
Verbo
SE-Passivo


Adjunto
Modo
Direção
Companhia
Adj_Causal
Comparação
Exemplo
Situação
Discurso
Advérbio