

# Recovering Capitalization and Punctuation Marks on Speech Transcriptions

Plan, structure, development

Options, difficulties and good experiments

---

Fernando Batista

# Recovering Capitalization and Punctuation Marks on Speech Transcriptions

Plan, structure, development

Options, difficulties and good experiments

---

Fernando Batista

## Contents

- Introduction
- Initial Plan
- Research overview
- Writing the research document
- Difficulties and good experiences
- Some advices

# Introduction

---

- Large quantities of audio and video being broadcasted daily
  - Automatic Speech Recognition (ASR) can provide additional information
    - Important for indexing, retrieval, subtitling, ...
- Audimus
  - A speech processing system available for Portuguese
  - Developed at L2F - INESC-ID



# Introduction

---

- The Automatic Speech Recognition (ASR) transcript is usually difficult to read and to process
  - Lack of punctuation/ inappropriate segmentation
  - Lack of capitalization
  - Disfluencies
  - Recognition errors



# Introduction

---

- The Automatic Speech Recognition (ASR) transcript is usually difficult to read and to process
  - Lack of punctuation/ inappropriate segmentation
  - Lack of capitalization
  - Disfluencies
  - Recognition errors
- Make the transcripts more readable (rich transcript)
  - Recover Punctuation Marks
  - Recover Capitalization
  - ...



# Exemplo:

## Texto extraído de um reconhecedor de fala

---

boa tarde a ministra da educação pronunciou sobre a polémica do professor suspenso  
maria de lurdes rodrigues disse que vai aguardar pelos resultados do processo que está a  
decorrer  
e garantiu que não tem motivos para duvidar do funcionamento da direcção regional de  
educação do norte  
que suspendeu passou por ter feito um comentário pose do primeiro-ministro  
a ministra disse lamentar que este tipo de pesados marquem  
à agenda mediática  
até este momento do muito  
que o li e ouvi  
não tenho nenhum sinal  
não tenho nenhum motivo para duvidar do funcionamento das instituições  
ou para de a considerar que pode estar em causa o funcionamento da direcção regional  
...

# Exemplo:

## Introdução de pontuação e segmentação

---

boa tarde.

a ministra da educação pronunciou sobre a polémica do professor suspenso

maria de lurdes rodrigues disse que vai aguardar pelos resultados do processo que está a decorrer

e garantiu que não tem motivos para duvidar do funcionamento da direcção regional de educação do norte

que suspendeu passou por ter feito um comentário pose do primeiro-ministro

a ministra disse lamentar que este tipo de pesados marquem

à agenda mediática

até este momento do muito

que o li e ouvi

não tenho nenhum sinal

não tenho nenhum motivo para duvidar do funcionamento das instituições

...

# Exemplo:

## Introdução de pontuação e segmentação

---

boa tarde.

a ministra da educação pronunciou sobre a polémica do professor suspenso.

maria de lurdes rodrigues disse que vai aguardar pelos resultados do processo que está a decorrer e garantiu que não tem motivos para duvidar do funcionamento da direcção regional de educação do norte

que suspendeu passou por ter feito um comentário pose do primeiro-ministro

a ministra disse lamentar que este tipo de pesados marquem

à agenda mediática

até este momento do muito

que o li e ouvi

não tenho nenhum sinal

não tenho nenhum motivo para duvidar do funcionamento das instituições

ou para de a considerar que pode estar em causa o funcionamento da direcção regional

...



# Exemplo:

## Introdução de pontuação e segmentação

---

boa tarde.

a ministra da educação pronunciou sobre a polémica do professor suspenso.

maria de lurdes rodrigues disse que vai aguardar pelos resultados do processo que está a decorrer e garantiu que não tem motivos para duvidar do funcionamento da direcção regional de educação do norte, que suspendeu passou por ter feito um comentário pose do primeiro-ministro.

a ministra disse lamentar que este tipo de pesados marquem

à agenda mediática

até este momento do muito

que o li e ouvi

não tenho nenhum sinal

não tenho nenhum motivo para duvidar do funcionamento das instituições

ou para de a considerar que pode estar em causa o funcionamento da direcção regional de educação dos seus de serviços

...

# Exemplo:

## Introdução de pontuação e segmentação

---

boa tarde.

a ministra da educação pronunciou sobre a polémica do professor suspenso.

maria de lurdes rodrigues disse que vai aguardar pelos resultados do processo que está a decorrer e garantiu que não tem motivos para duvidar do funcionamento da direcção regional de educação do norte, que suspendeu passou por ter feito um comentário pose do primeiro-ministro.

a ministra disse lamentar que este tipo de pesados marquem à agenda mediática.

até este momento do muito

que o li e ouvi

não tenho nenhum sinal

não tenho nenhum motivo para duvidar do funcionamento das instituições

ou para de a considerar que pode estar em causa o funcionamento da direcção regional de educação dos seus de serviços

aquilo que é a minha preocupação é que no âmbito deste processo estejam garantidos

...

# Exemplo:

## Introdução de pontuação e segmentação

---

boa tarde.

a ministra da educação pronunciou sobre a polémica do professor suspenso.

maria de lurdes rodrigues disse que vai aguardar pelos resultados do processo que está a decorrer e garantiu que não tem motivos para duvidar do funcionamento da direcção regional de educação do norte, que suspendeu passou por ter feito um comentário pose do primeiro-ministro.

a ministra disse lamentar que este tipo de pesados marquem à agenda mediática.

até este momento do muito que o li e ouvi

não tenho nenhum sinal

não tenho nenhum motivo para duvidar do funcionamento das instituições

ou para de a considerar que pode estar em causa o funcionamento da direcção regional

de educação dos seus de serviços

aquilo que é a minha preocupação é que no âmbito deste processo estejam garantidos

a existência dos mecanismos de defesa

# Exemplo:

## Introdução de pontuação e segmentação

---

boa tarde.

a ministra da educação pronunciou sobre a polémica do professor suspenso.

maria de lurdes rodrigues disse que vai aguardar pelos resultados do processo que está a decorrer e garantiu que não tem motivos para duvidar do funcionamento da direcção regional de educação do norte, que suspendeu passou por ter feito um comentário pose do primeiro-ministro.

a ministra disse lamentar que este tipo de pesados marquem à agenda mediática.

até este momento do muito que o li e ouvi não tenho nenhum sinal.

não tenho nenhum motivo para duvidar do funcionamento das instituições

ou para de a considerar que pode estar em causa o funcionamento da direcção regional de educação dos seus de serviços

aquilo que é a minha preocupação é que no âmbito deste processo estejam garantidos a existência dos mecanismos de defesa

# Exemplo:

## Introdução de pontuação e segmentação

---

boa tarde.

a ministra da educação pronunciou sobre a polémica do professor suspenso.

maria de lurdes rodrigues disse que vai aguardar pelos resultados do processo que está a decorrer e garantiu que não tem motivos para duvidar do funcionamento da direcção regional de educação do norte, que suspendeu passou por ter feito um comentário pose do primeiro-ministro.

a ministra disse lamentar que este tipo de pesados marquem à agenda mediática.

até este momento do muito que o li e ouvi não tenho nenhum sinal.

não tenho nenhum motivo para duvidar do funcionamento das instituições ou para de a considerar que pode estar em causa o funcionamento da direcção regional

de educação dos seus de serviços

aquilo que é a minha preocupação é que no âmbito deste processo estejam garantidos

a existência dos mecanismos de defesa

# Exemplo:

## Introdução de pontuação e segmentação

---

boa tarde.

a ministra da educação pronunciou sobre a polémica do professor suspenso.

maria de lurdes rodrigues disse que vai aguardar pelos resultados do processo que está a decorrer e garantiu que não tem motivos para duvidar do funcionamento da direcção regional de educação do norte, que suspendeu passou por ter feito um comentário pose do primeiro-ministro.

a ministra disse lamentar que este tipo de pesados marquem à agenda mediática.

até este momento do muito que o li e ouvi não tenho nenhum sinal.

não tenho nenhum motivo para duvidar do funcionamento das instituições ou para de a considerar que pode estar em causa o funcionamento da direcção regional de educação, dos seus de serviços.

aquilo que é a minha preocupação é que no âmbito deste processo estejam garantidos

a existência dos mecanismos de defesa

# Exemplo:

## Introdução de pontuação e segmentação

---

boa tarde.

a ministra da educação pronunciou sobre a polémica do professor suspenso.

maria de lurdes rodrigues disse que vai aguardar pelos resultados do processo que está a decorrer e garantiu que não tem motivos para duvidar do funcionamento da direcção regional de educação do norte, que suspendeu passou por ter feito um comentário pose do primeiro-ministro.

a ministra disse lamentar que este tipo de pesados marquem à agenda mediática.

até este momento do muito que o li e ouvi não tenho nenhum sinal.

não tenho nenhum motivo para duvidar do funcionamento das instituições ou para de a considerar que pode estar em causa o funcionamento da direcção regional de educação, dos seus de serviços.

aquilo que é a minha preocupação é que no âmbito deste processo estejam garantidos a existência dos mecanismos de defesa.

# Exemplo:

## Melhor ainda se tiver maiúsculas

---

Boa tarde.

A ministra da **Educação** pronunciou sobre a polémica do professor suspenso.

**Maria de Lurdes Rodrigues** disse que vai aguardar pelos resultados do processo que está a decorrer e garantiu que não tem motivos para duvidar do funcionamento da **Direcção Regional de Educação do Norte**, que suspendeu passou por ter feito um comentário pose do primeiro-ministro.

A ministra disse lamentar que este tipo de pesados marquem à agenda mediática.

**Até** este momento do muito que o li e ouvi não tenho nenhum sinal.

**Não** tenho nenhum motivo para duvidar do funcionamento das instituições ou para de a considerar que pode estar em causa o funcionamento da **Direcção Regional de Educação**, dos seus de serviços.

**Aquilo** que é a minha preocupação é que no âmbito deste processo estejam garantidos a existência dos mecanismos de defesa.



# Exemplo:

## Cuidado com os erros de reconhecimento

---

Boa tarde.

A ministra da Educação pronunciou sobre a polémica do professor suspenso.

Maria de Lurdes Rodrigues disse que vai aguardar pelos resultados do processo que está a decorrer e garantiu que não tem motivos para duvidar do funcionamento da Direcção Regional de Educação do Norte, que suspendeu ~~passou~~ por ter feito um comentário ~~pose~~ do primeiro-ministro.

A ministra disse lamentar que este tipo de ~~pesados~~ marquem à agenda mediática.

Até este momento do muito que ~~e~~ li e ouvi não tenho nenhum sinal.

Não tenho nenhum motivo para duvidar do funcionamento das instituições ou para ~~de-a~~ considerar que pode estar em causa o funcionamento da Direcção Regional de Educação, dos seus ~~de~~ serviços.

Aquilo que é a minha preocupação é que no âmbito deste processo estejam garantidos a existência dos mecanismos de defesa.

# Exemplo:

## Texto sem erros ...

---

Boa tarde.

A ministra da Educação pronunciou-se sobre a polémica do professor suspenso.

Maria de Lurdes Rodrigues disse que vai aguardar pelos resultados do processo que está a decorrer e garantiu que não tem motivos para duvidar do funcionamento da Direcção Regional de Educação do Norte, que suspendeu o professor por ter feito um comentário a propósito do primeiro-ministro.

A ministra disse lamentar que este tipo de episódios marquem a agenda mediática.

Até este momento do muito que li e ouvi não tenho nenhum sinal.

Não tenho nenhum motivo para duvidar do funcionamento das instituições ou para considerar que pode estar em causa o funcionamento da Direcção Regional de Educação, ou dos seus serviços.

Aquilo que é a minha preocupação é que no âmbito deste processo estejam garantidos a existência dos mecanismos de defesa.

# Exemplo:

## ... ainda assim diferente de um texto escrito...

---

Boa tarde.

A ministra da Educação pronunciou-se sobre a polémica do professor suspenso.

Maria de Lurdes Rodrigues disse que vai aguardar pelos resultados do processo que está a decorrer e garantiu que não tem motivos para duvidar do funcionamento da Direcção Regional de Educação do Norte, que suspendeu o professor por ter feito um comentário a propósito do primeiro-ministro.

A ministra disse lamentar que este tipo de episódios marquem a agenda mediática.

Até este momento do muito que li e ouvi não tenho nenhum sinal.

Não tenho nenhum motivo para duvidar do funcionamento das instituições {BREATH} ou para %aa considerar que pode estar em causa o funcionamento da Direcção Regional de Educação, ou dos seus %aa serviços.

Aquilo que é a minha preocupação é que no âmbito deste processo estejam garantidos a existência dos mecanismos de de defesa.

# Initial Plan

---

## Enriquecimento de transcrições com recurso a

### Técnicas de processamento de língua natural escrita

Dada a transcrição de fala espontânea, tal como uma reunião ou uma aula, o objectivo é desenvolver métodos que permitam produzir automaticamente um novo nível de transcrição que envolve:

- (?) *Camada de meta-informação proveniente do conhecimento da língua natural escrita*
- Correções
- **Separação das frases**
- Tratamento da pontuação (vírgulas, ...)
- **Capitalização** (Nomes, Inícios de frase, ...)
- Identificação de frases interrogativas (pontuação)
- **Identificação de datas, dinheiro, números** (cardinais e ordinais)
- Na fala espontânea ocorrem também com alguma frequência fenómenos da linguagem que, quando transcritos literalmente, tornam a leitura difícil, como é o caso de:
  - Pausas preenchidas.
  - **Outras disfluências** de linguagem, que devem ser marcadas
    - hesitações
    - **repetição de palavras.**

Este novo nível de transcrição enriquecida deverá ser:

- **Mais inteligível** - serão efectuadas correções com base no conhecimento da estrutura da língua.
- *mais adequado para o processamento computacional dado o conjunto de meta-informação introduzida*

### Estratégia

- **Analisar o resultado** proveniente do **reconhecedor**. Alinhar com transcrição manual para:
  - Levantamento de fenómenos particulares que interessa analisar.
  - Identificar a frequência e distribuição de cada fenómeno particular
- Proceder a uma recolha bibliográfica das formas actualmente utilizadas na detecção automática de disfluências em transcrições automáticas.
- Elaborar um estudo da integração da análise sintáctica com a informação prosódica, para inserção automática de marcas de pontuação.

### Caminhos alternativos

- Considerar informação proveniente do sinal de fala, tal como a prosódia (?)
- Analisar a possibilidade de utilização de grafos de palavras provenientes do reconhecimento de fala (lattices), contendo as diferentes alternativas de transcrição.

# Initial Plan

---

## Enriquecimento de transcrições com recurso a

### Técnicas de processamento de língua natural escrita

Dada a transcrição de fala espontânea, tal como uma reunião ou uma aula, o objectivo é desenvolver métodos que permitam produzir automaticamente um novo nível de transcrição que envolva:

- *(?) Camada de meta-informação proveniente do conhecimento da língua natural escrita*
- Correções
- **Separação das frases**
- Tratamento da pontuação (vírgulas, ...)
- **Capitalização** (Nomes, Inícios de frase, ...)
- Identificação de frases interrogativas (pontuação)
- **Identificação de datas, dinheiro, números** (cardinais e ordinais)
- Na fala espontânea ocorrem também com alguma frequência fenómenos da linguagem que, quando transcritos literalmente, tornam a leitura difícil, como é o caso de:
  - Pausas preenchidas.
  - **Outras disfluências** de linguagem, que devem ser marcadas
    - hesitações

# Initial Plan

- Na fala espontânea ocorrem também com alguma frequência fenómenos da linguagem que, quando transcritos literalmente, tornam a leitura difícil, como é o caso de:
  - Pausas preenchidas.
  - **Outras disfluências** de linguagem, que devem ser marcadas
    - hesitações
    - **repetição de palavras.**

Este novo nível de transcrição enriquecida deverá ser:

- **Mais inteligível** - serão efectuadas correcções com base no conhecimento da estrutura da língua.
- *mais adequado para o processamento computacional dado o conjunto de meta-informação introduzida*

## Estratégia

- **Analisar o resultado** proveniente do **reconhecedor**. Alinhar com transcrição manual para:
  - Levantamento de fenómenos particulares que interessa analisar.
  - Identificar a frequência e distribuição de cada fenómeno particular
- Proceder a uma recolha bibliográfica das formas actualmente utilizadas na

# Initial Plan

## Estratégia

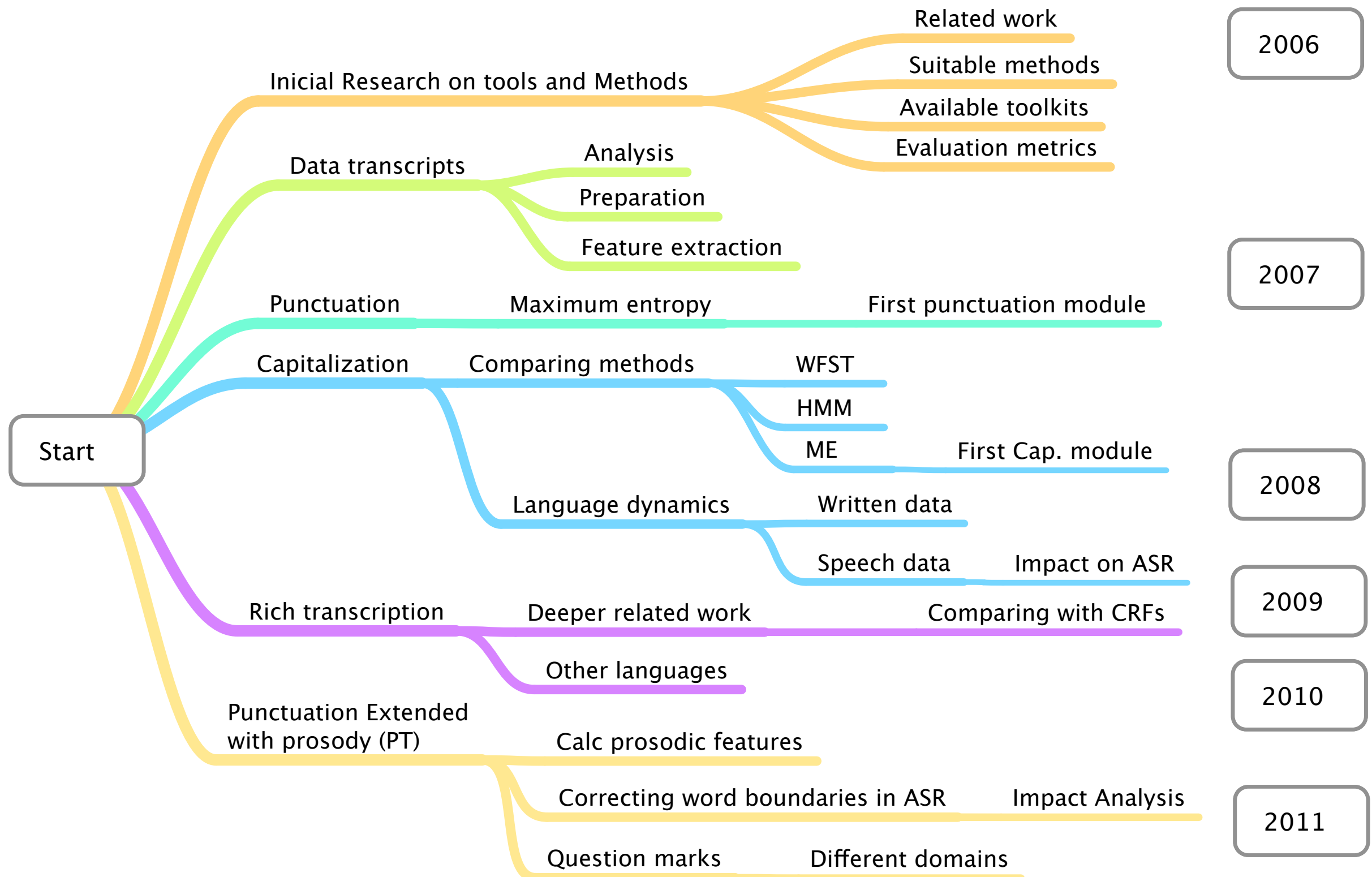
- **Analisar o resultado** proveniente do **reconhecedor**. Alinhar com transcrição manual para:
  - Levantamento de fenómenos particulares que interessa analisar.
  - Identificar a frequência e distribuição de cada fenómeno particular
- Proceder a uma recolha bibliográfica das formas actualmente utilizadas na detecção automática de disfluências em transcrições automáticas.
- Elaborar um estudo da integração da análise sintáctica com a informação prosódica, para inserção automática de marcas de pontuação.

## Caminhos alternativos

- Considerar informação proveniente do sinal de fala, tal como a prosódia (?)
- Analisar a possibilidade de utilização de grafos de palavras provenientes do reconhecimento de fala (lattices), contendo as diferentes alternativas de transcrição.

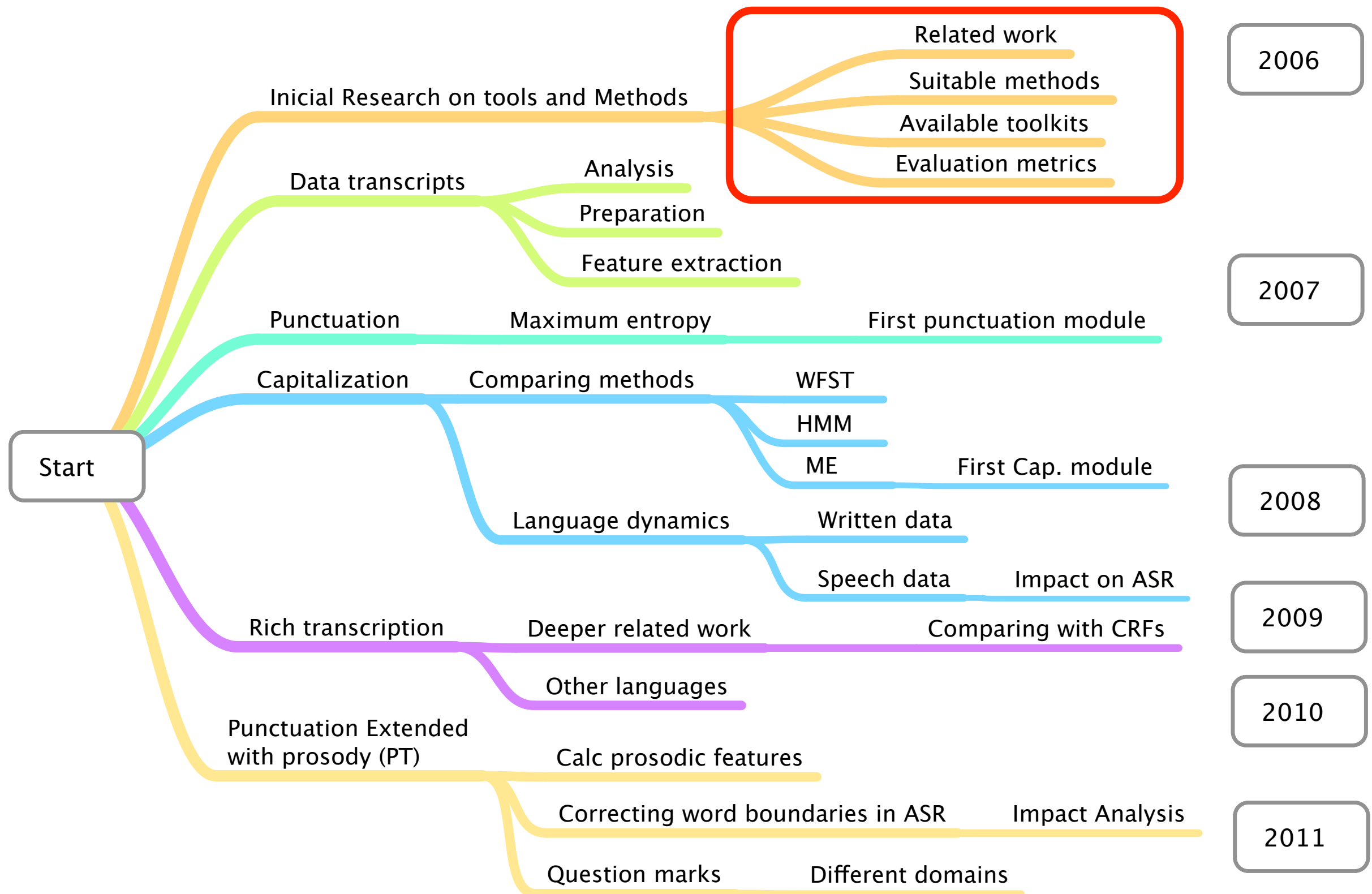
# Work Overview

demo demo

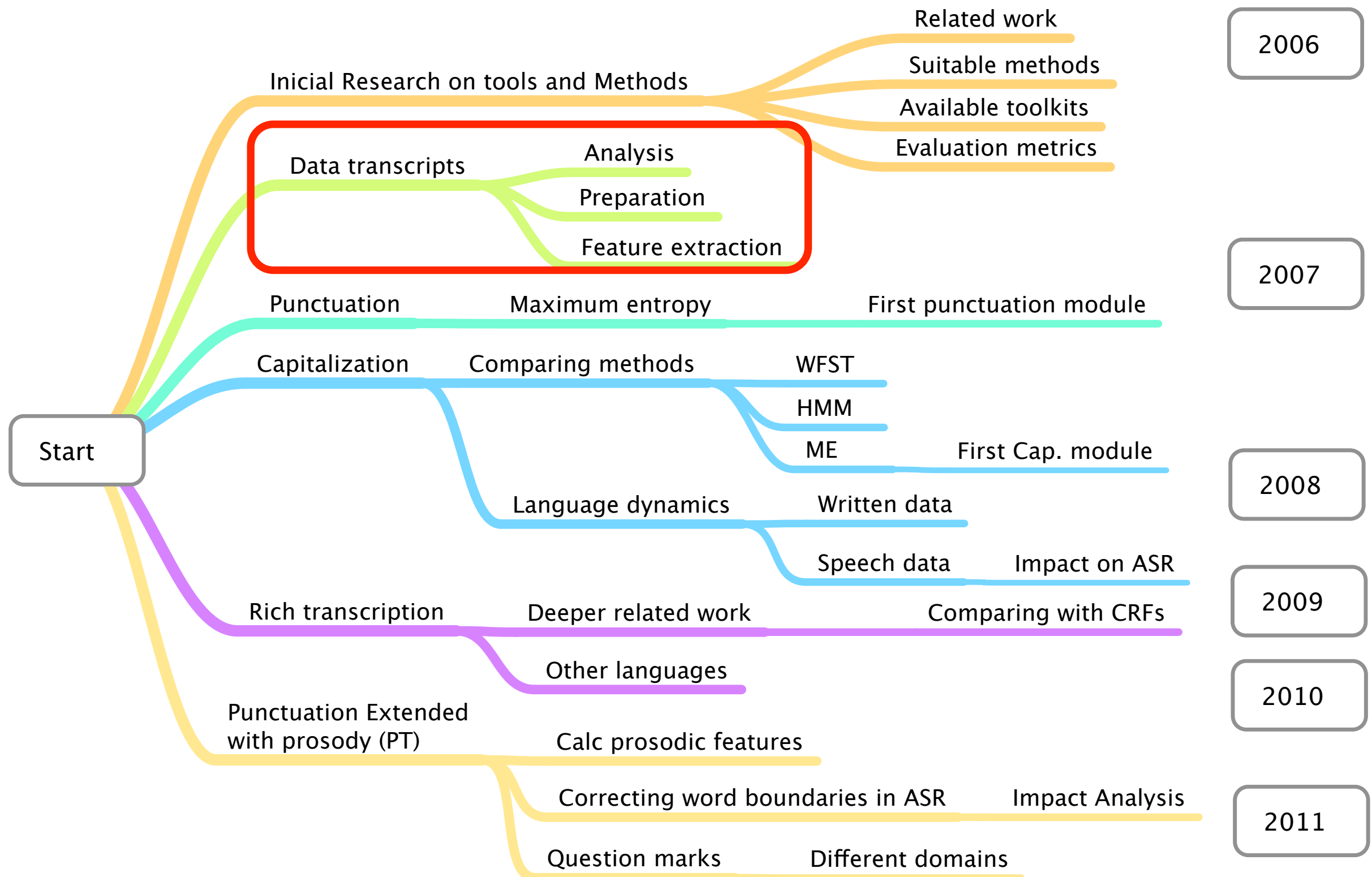




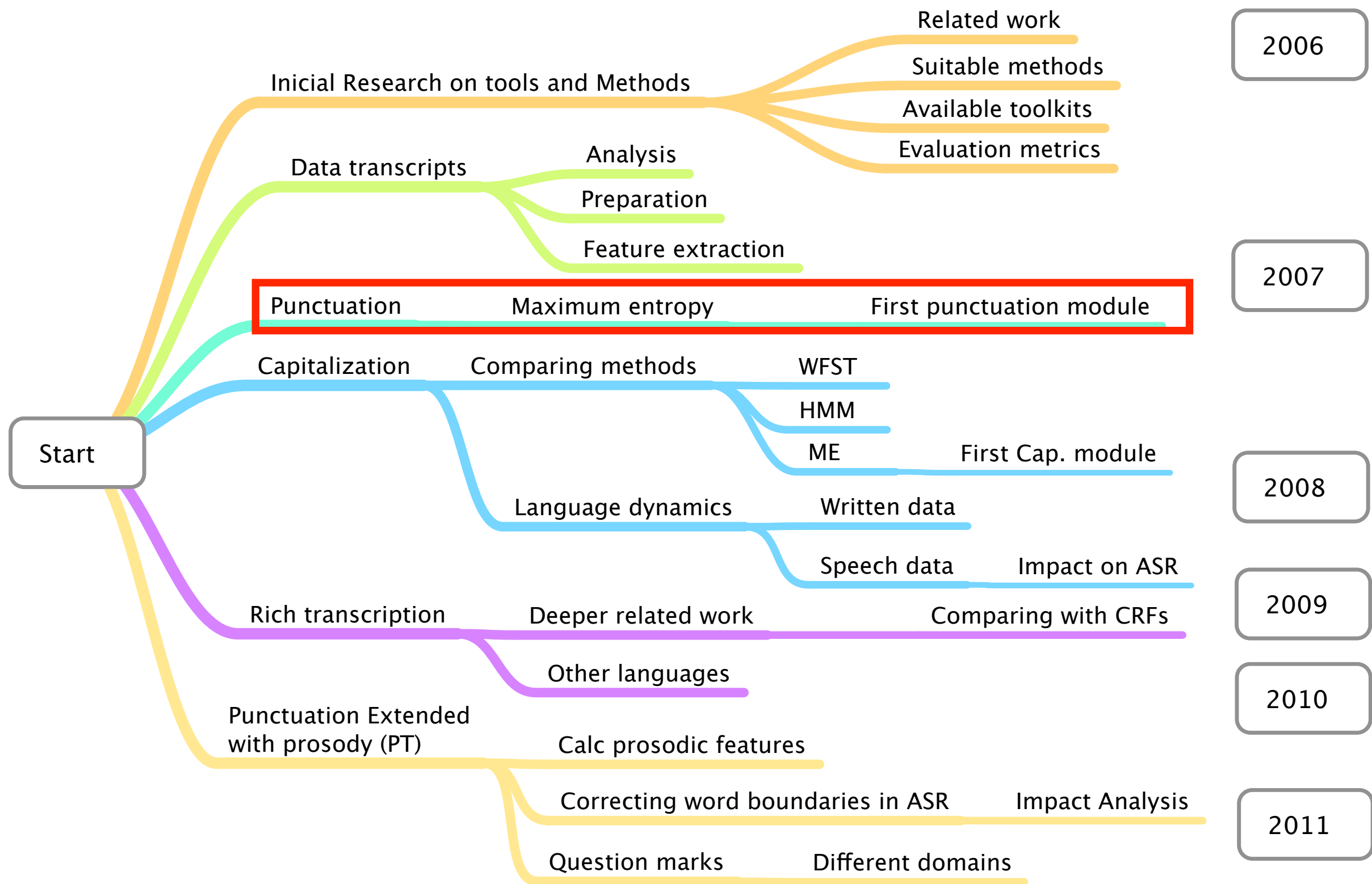
# Work Overview



# Work Overview

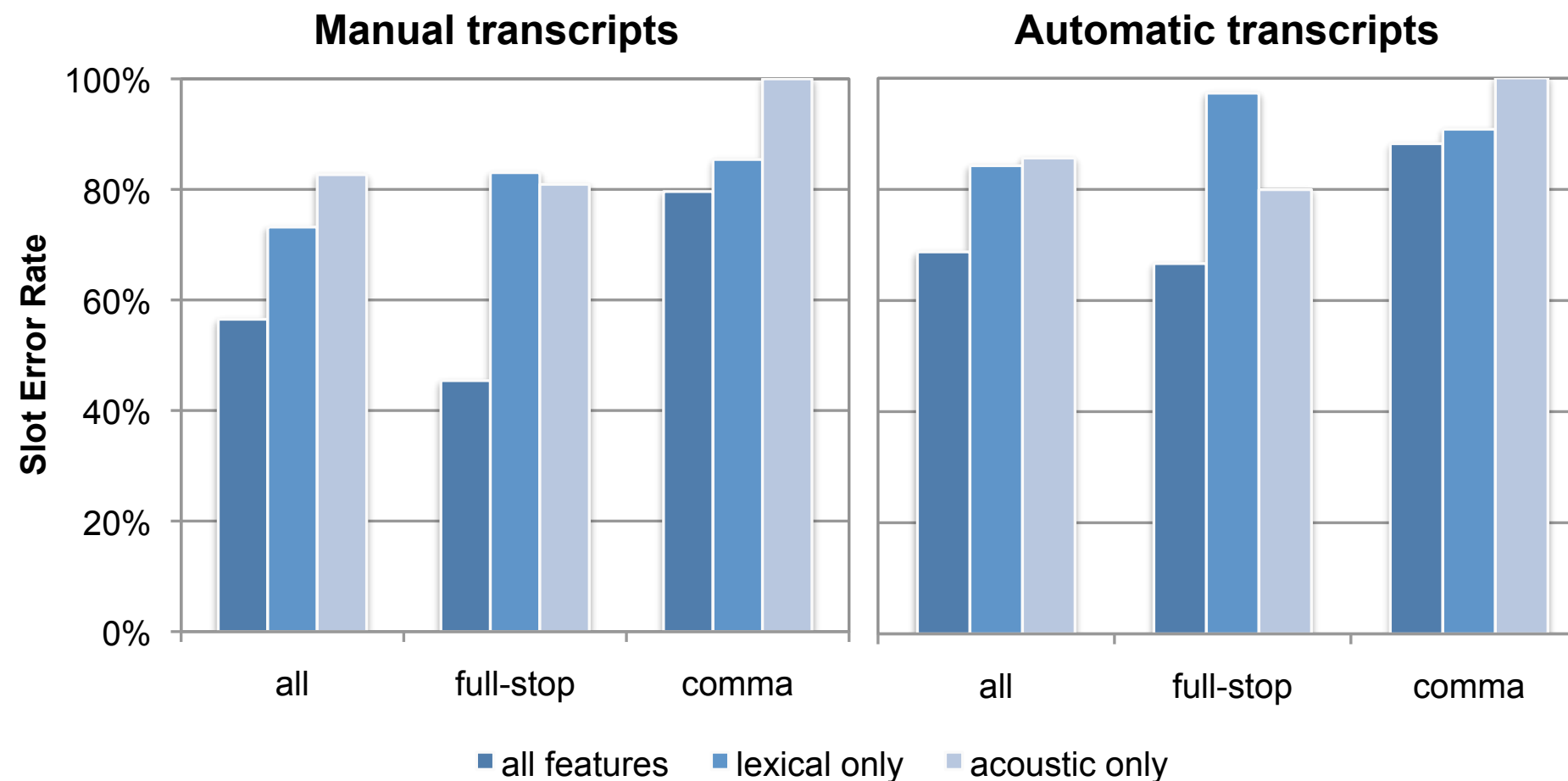


# Work Overview



# Recovering full-stop and comma

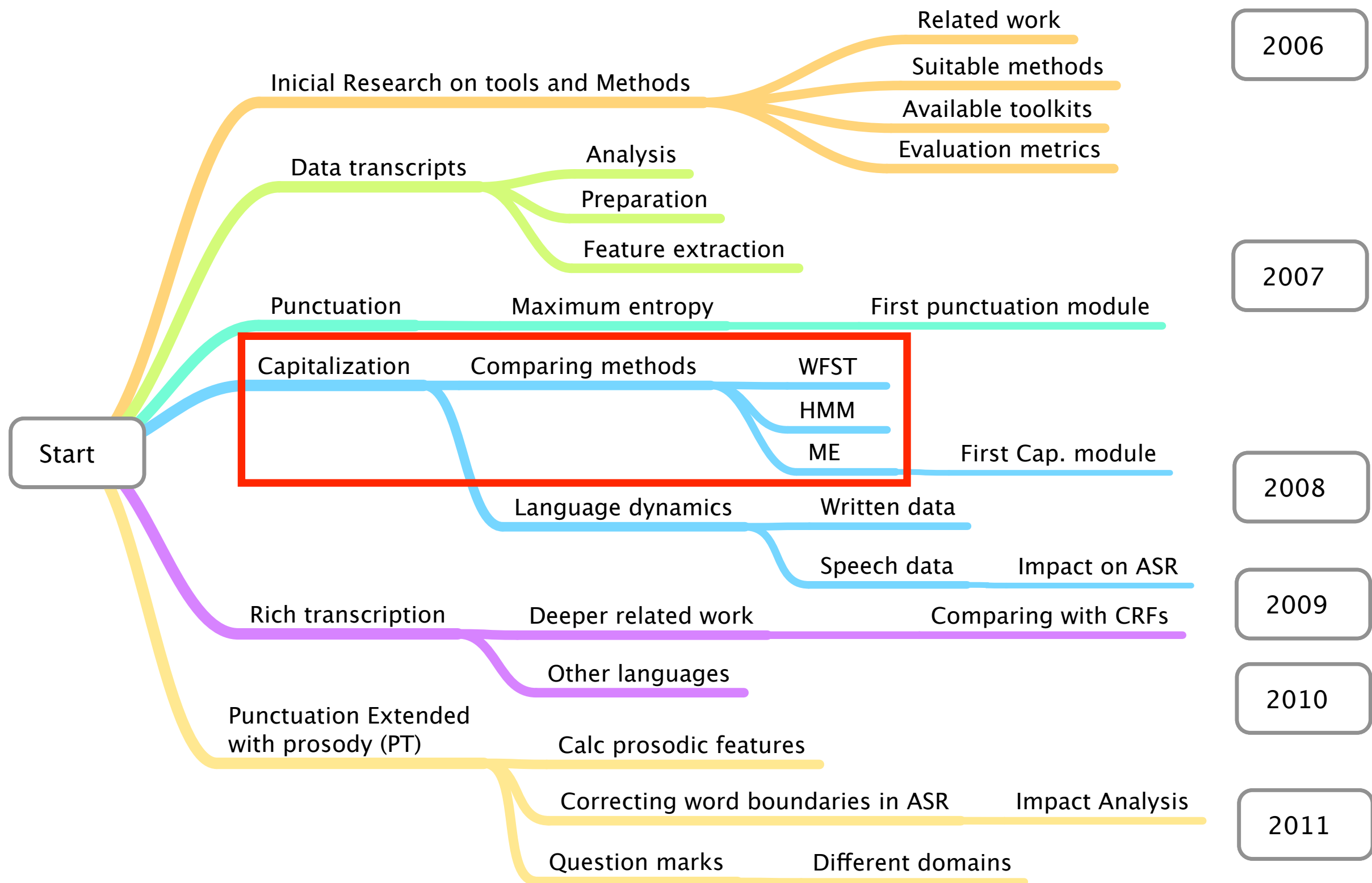
## Feature analysis



- **Conclusions**

- Larger impact of lexical features in general, but more in the *comma* detection
- Larger impact of acoustic features in the *full-stop* detection
- Combining all the information leads to the best results

# Work Overview



- **HMM-based tagger (SRILM toolkit) (Stolcke, 2002)**
  - Often used for this task
  - Models the capitalization context well
  - Generative training
- **WFST-based**
- **Maximum Entropy (ME) models**
  - Also used for punctuation recovery
  - Allows a richer set of features
  - Discriminative training

# Capitalization Task

## Early work comparing approaches



technology  
from seed

- **HMM-based tagger (SRILM toolkit) (Stolcke, 2002)**

- Often used for this task
- Models the capitalization context well
- Generative training

- **WFST-based**

**Better for written corpora**  
captures well the written corpora structure

- **Maximum Entropy (ME) models**

- Also used for punctuation recovery
- Allows a richer set of features
- Discriminative training

# Capitalization Task

## Early work comparing approaches



technology  
from seed

- **HMM-based tagger (SRILM toolkit)** (Stolcke, 2002)
  - Often used for this task
  - Models the capitalization context well
  - Generative training

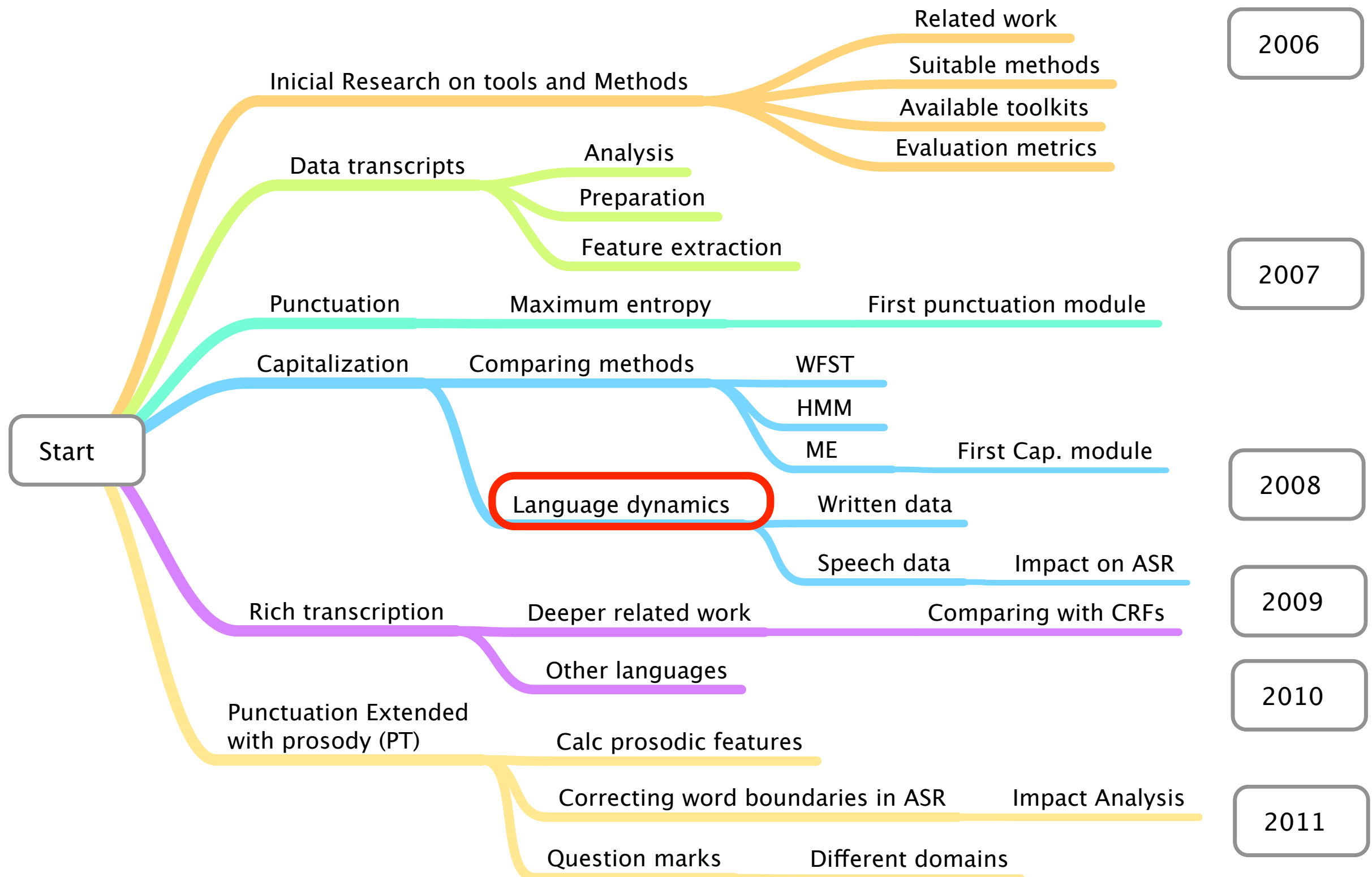
**Better for written corpora**  
captures well the written corpora structure

- **WFST-based**
- **Maximum Entropy (ME) models**
  - Also used for punctuation recovery
  - Allows a richer set of features
  - Discriminative training

**Better for speech transcripts**  
Include portions of **spontaneous speech**,  
with a **more flexible linguistic structure**  
when compared to written corpora



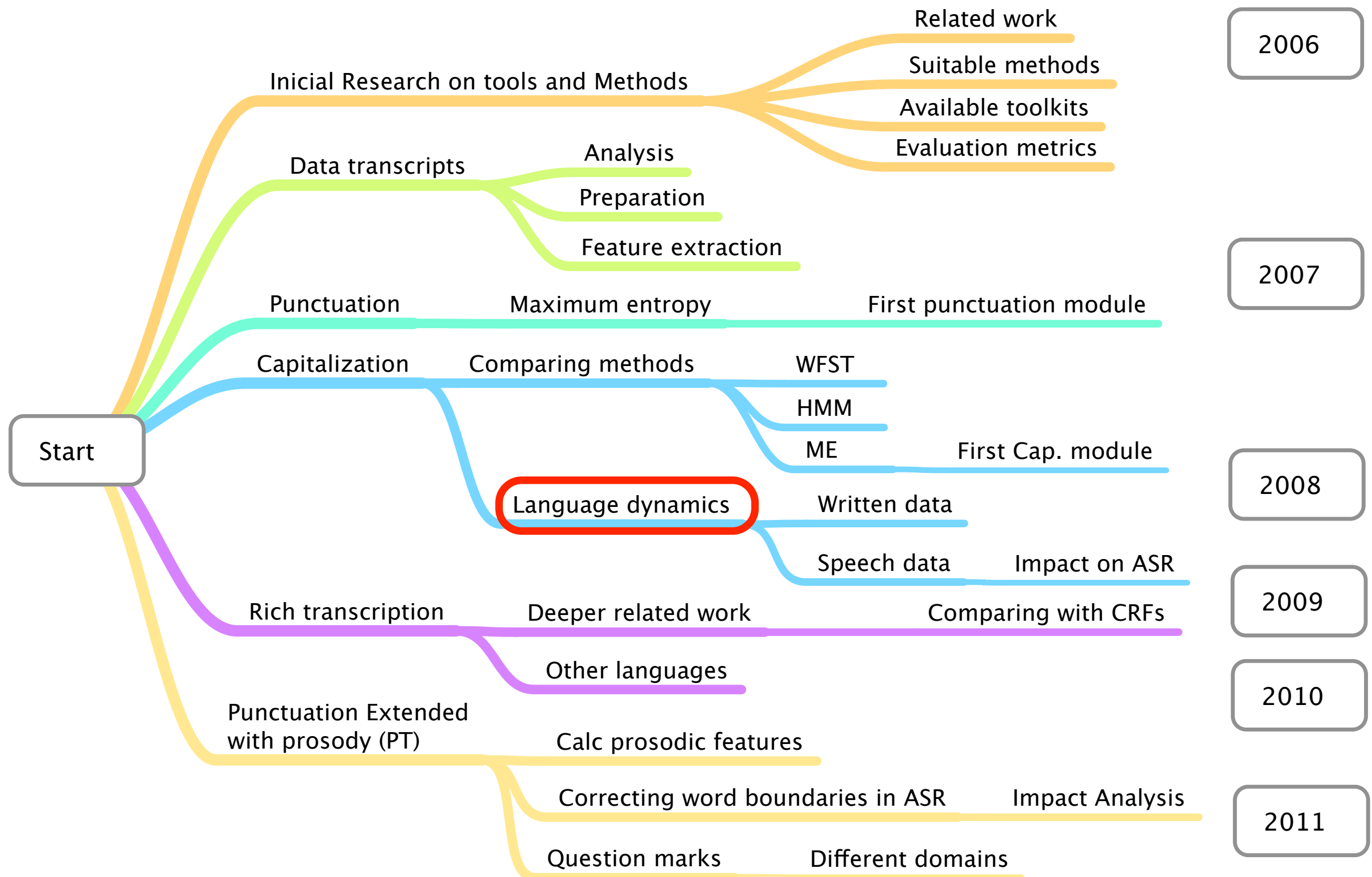
# Work Overview



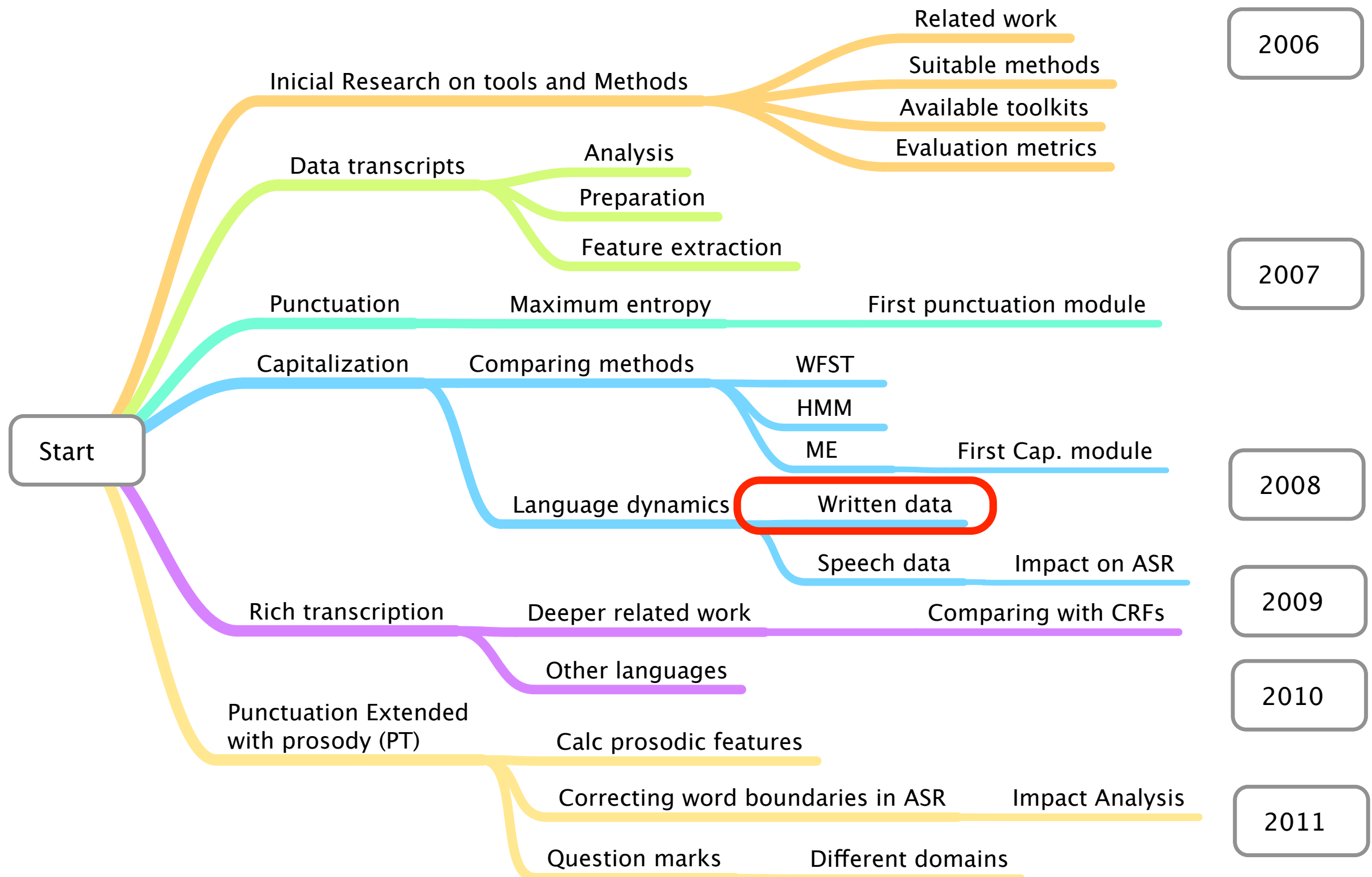
- **The language dynamics problem**
  - New words are introduced everyday
    - Brazil Elections: [Jair Bolsonaro](#), [Fernando Haddad](#)
    - Weather: [Super Typhoon Yutu](#), [Hardwick](#), ...
    - [Apu Nahasapeemapetilon](#)
    - [Staycation](#) ("vá para fora cá dentro")
  - The usage of other decays with time
    - Gulf war: [Khuzestan](#), [Abadan](#), [Khorramshahr](#), ...
    - Afghanistan: [CH-47 Chinook](#) helicopter, ...
    - Japan Tsunami: [Fukushima](#), [Honshu](#), [Oshika](#), ...
- **How does language dynamics affect the performance of the capitalization task?**  
**How to update the capitalization model?**



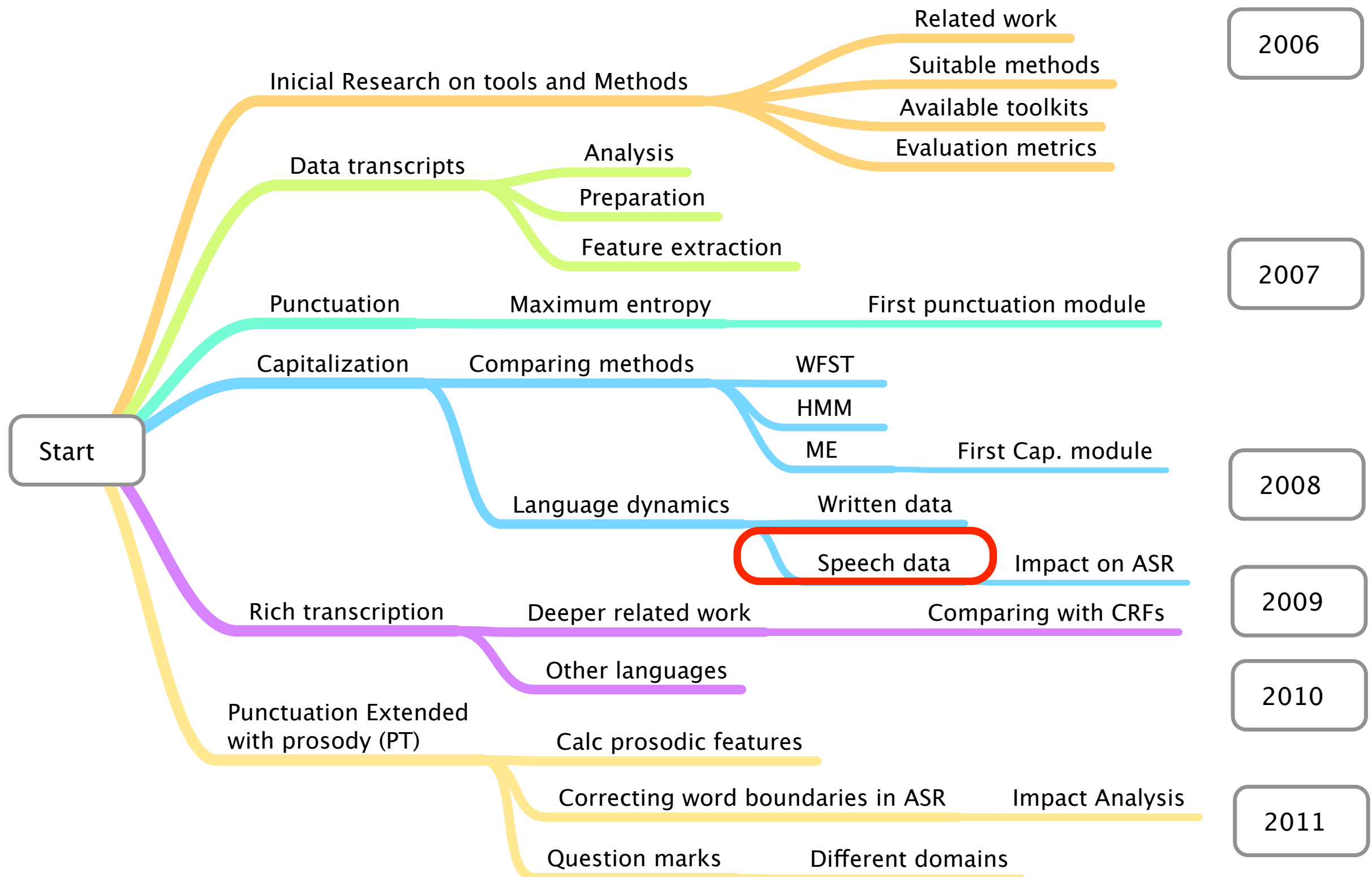
# Work Overview



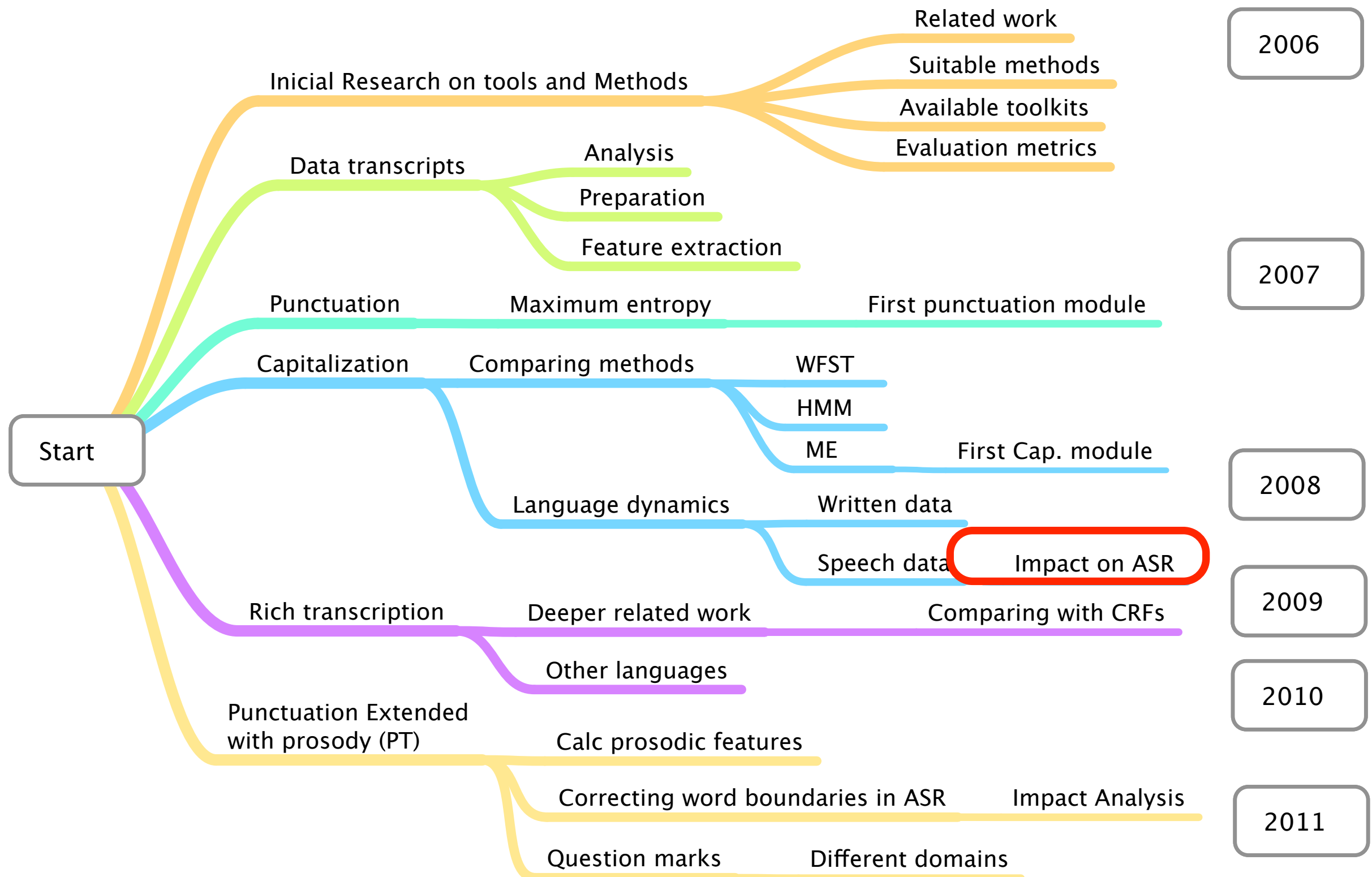
# Work Overview



# Work Overview



# Work Overview



# IMPACT OF DYNAMIC MODEL ADAPTATION BEYOND SPEECH RECOGNITION

*Fernando Batista<sup>1,2,3</sup>, Rui Amaral<sup>1,2,4</sup>, Isabel Trancoso<sup>1,2</sup>, Nuno Mamede<sup>1,2</sup>*

<sup>1</sup>*L<sup>2</sup>F* - Spoken Language Systems Laboratory - INESC ID Lisboa

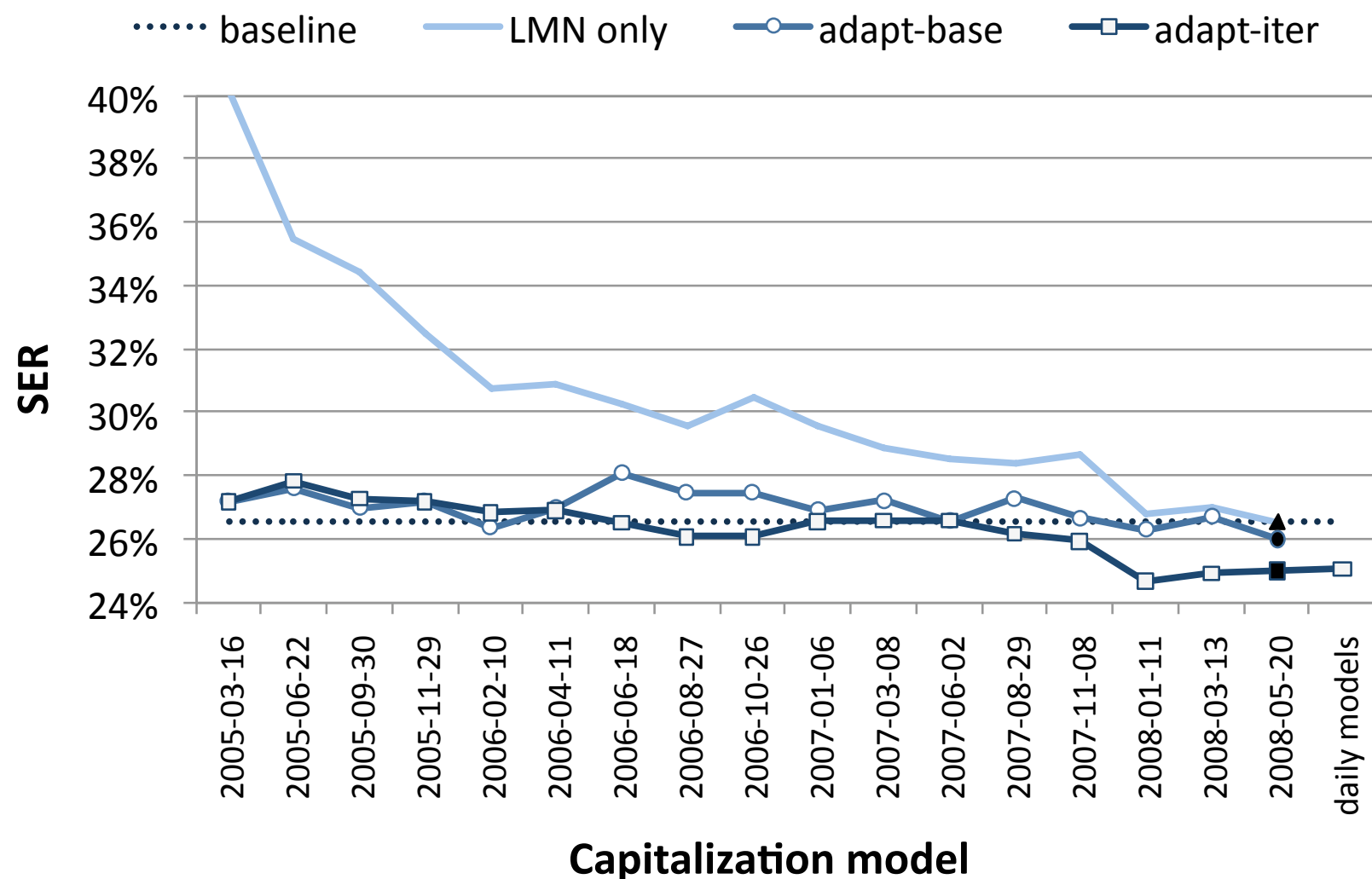
R. Alves Redol, 9, 1000-029 Lisboa, Portugal

<http://www.l2f.inesc-id.pt/>

<sup>2</sup>IST – Technical University of Lisbon, Portugal

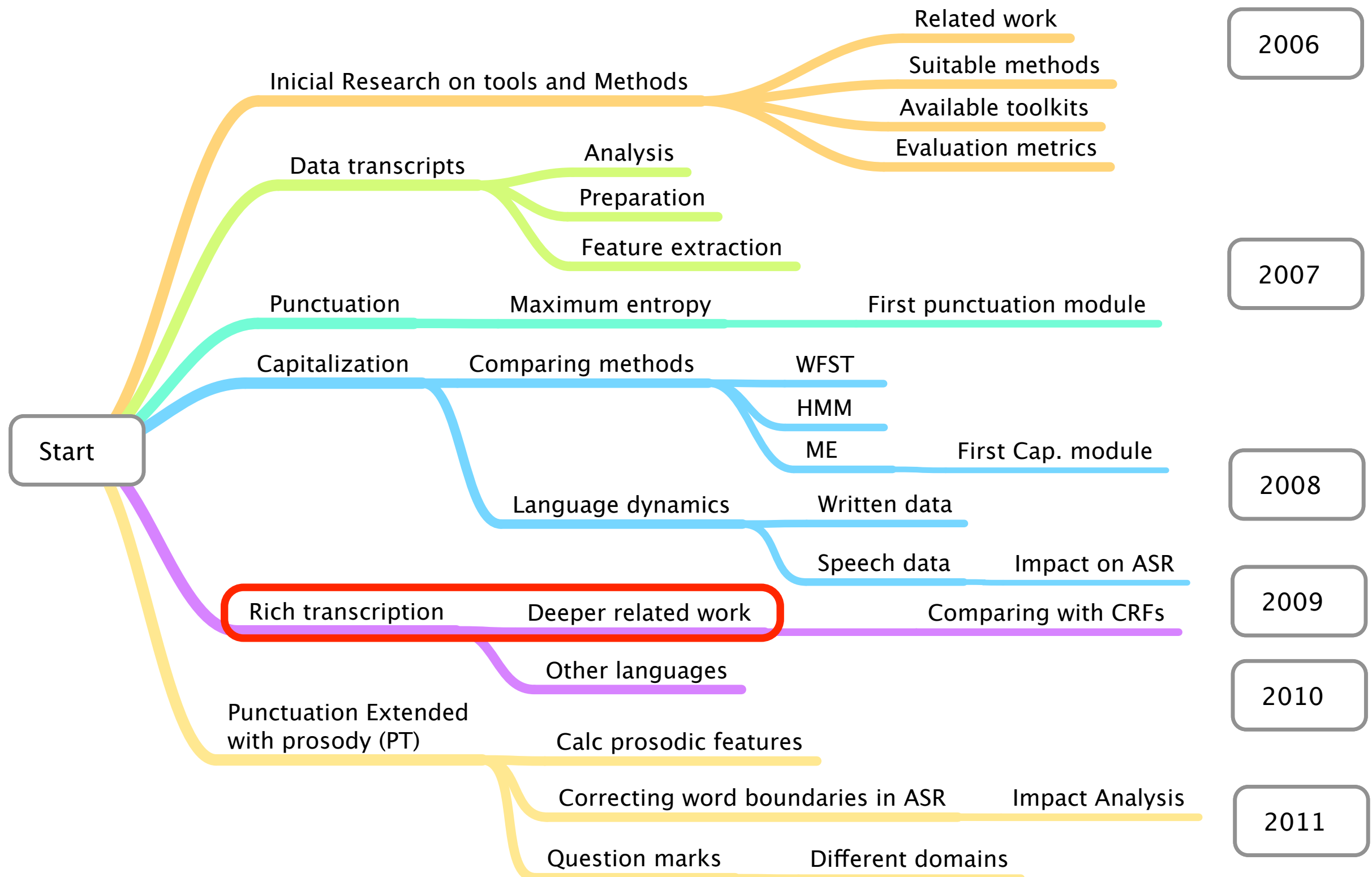
<sup>3</sup>ISCTE – Instituto de Ciências do Trabalho e da Empresa, Portugal

<sup>4</sup>EST – Escola Superior de Tecnologia de Setúbal



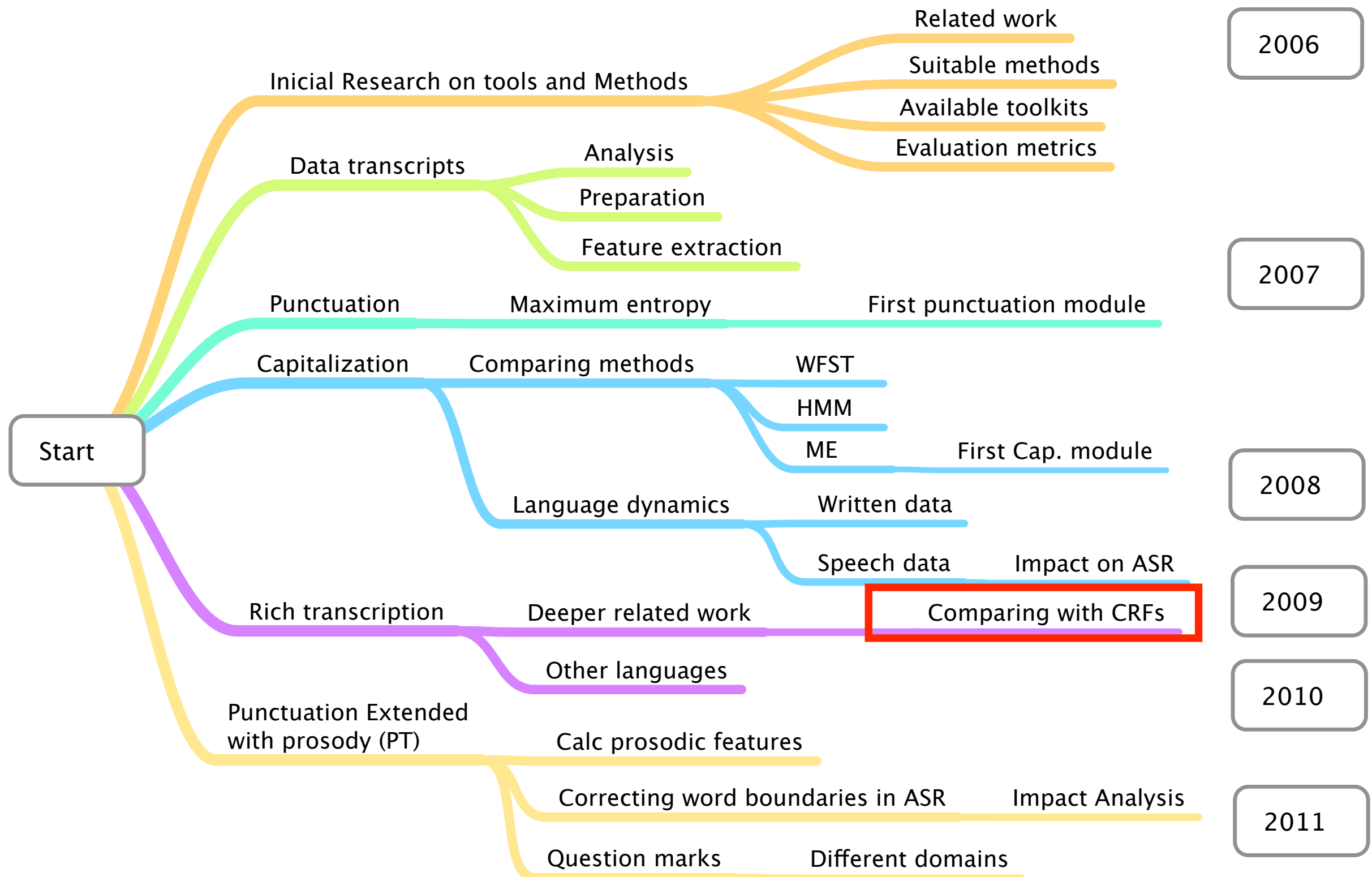


# Work Overview





# Work Overview



# Capitalization task

## Comparing all approaches



technology  
from seed

- Different approaches have been compared

- **HMM-based approach**

- Often used for this task
    - Models the capitalization context well
    - Generative training



**Better for written corpora**

captures well the written corpora structure

- **Maximum Entropy (ME) models**

- Also used for punctuation recovery
    - Allows a richer set of features
    - Discriminative training



**Suitable for speech transcripts**

Include portions of **spontaneous speech**,  
with a **more flexible linguistic structure**  
when compared to written corpora

- **CRF-based**

- Feature rich, like ME
    - Label dependency/context, like HMM

- Different approaches have been compared

- **HMM-based approach**

- Often used for this task
- Models the capitalization context well
- Generative training



**Better for written corpora**

captures well the written corpora structure

- **Maximum Entropy (ME) models**

- Also used for punctuation recovery
- Allows a richer set of features
- Discriminative training



**Suitable for speech transcripts**

Include portions of **spontaneous speech**, with a **more flexible linguistic structure** when compared to written corpora

- **CRF-based**

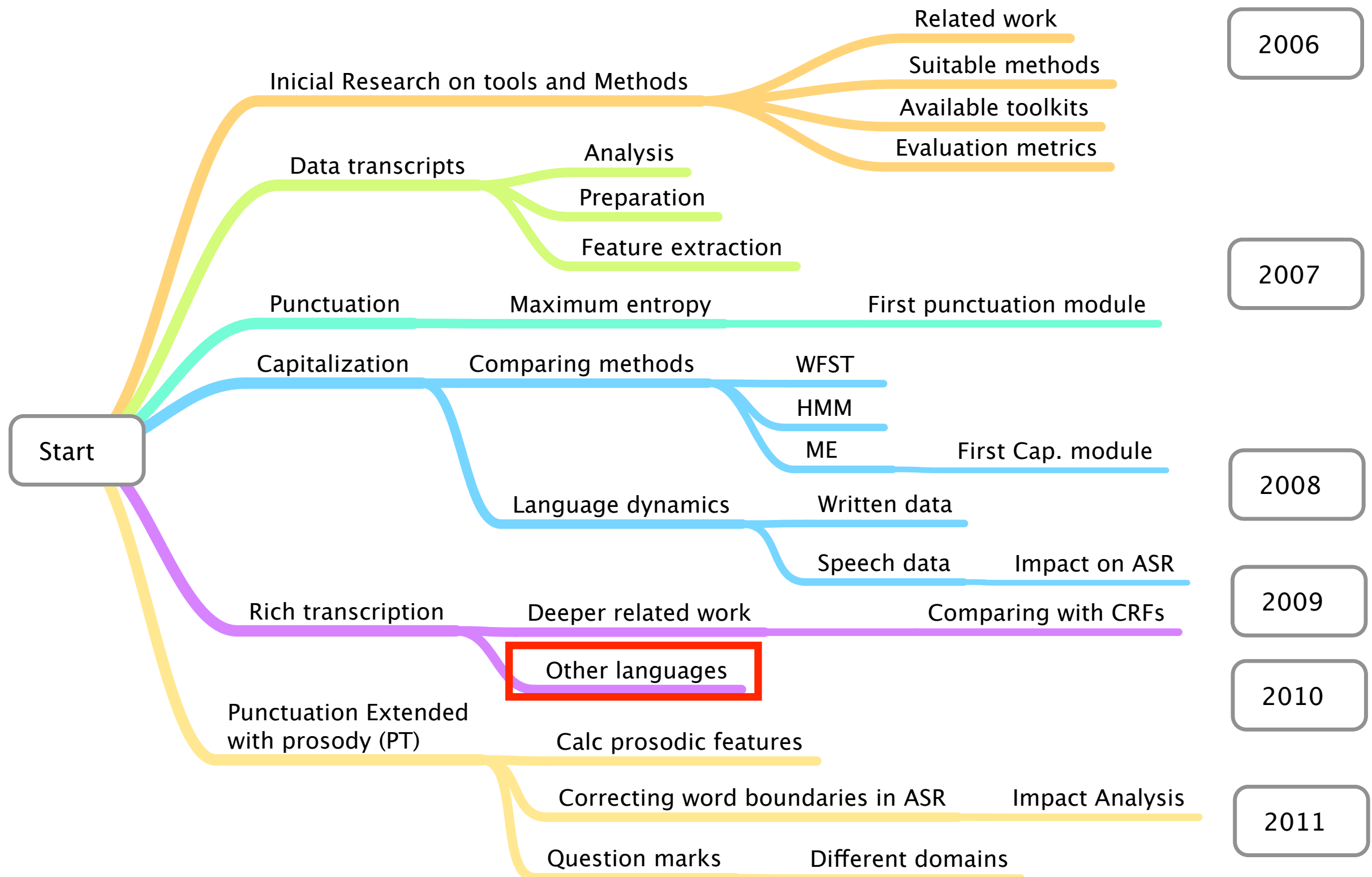
- Feature rich, like ME
- Label dependency/context, like HMM



**Even better than ME**

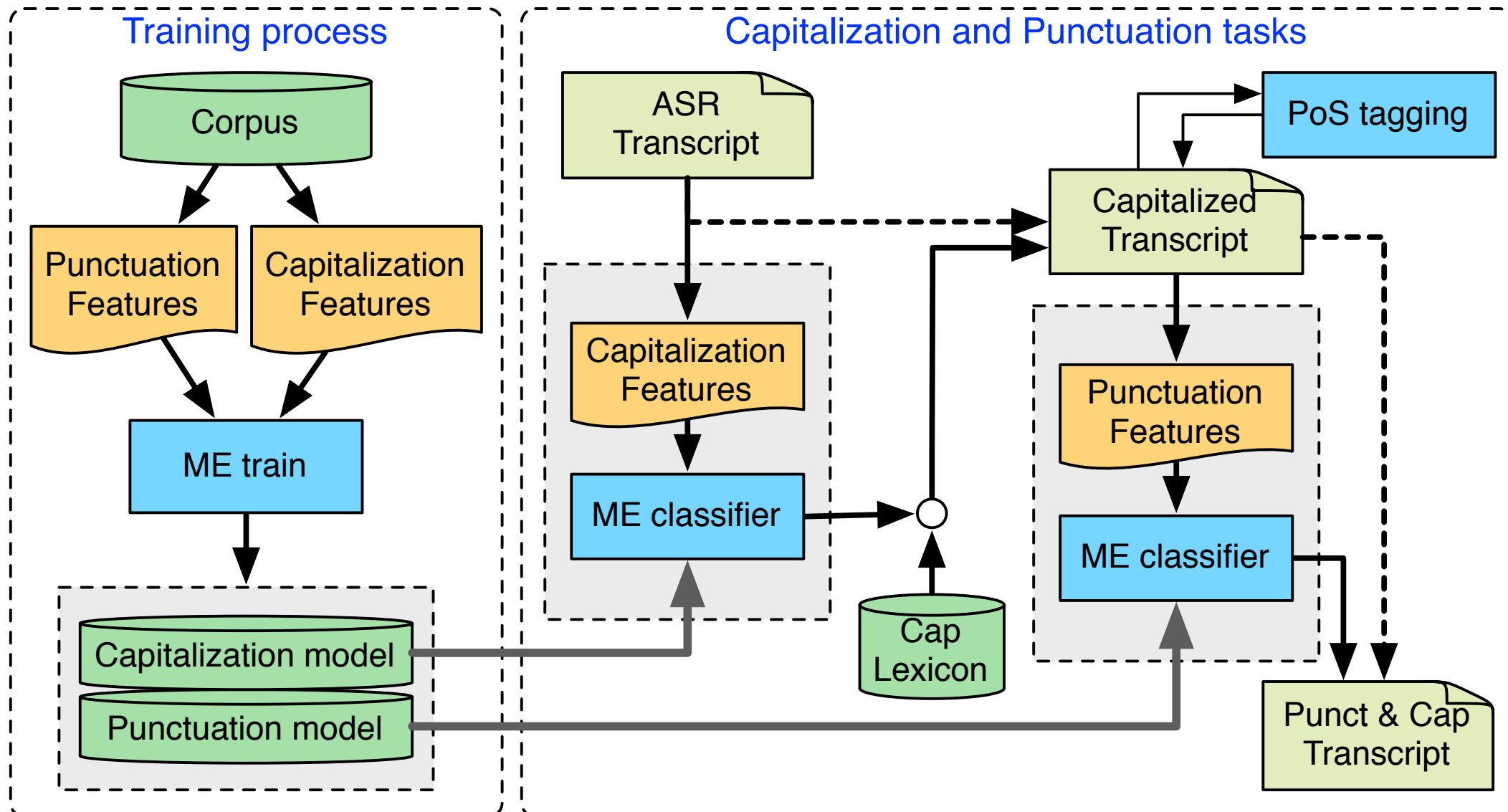
Supports the idea that the capitalization of a word tends to be connected with the capitalization of words around.

# Work Overview

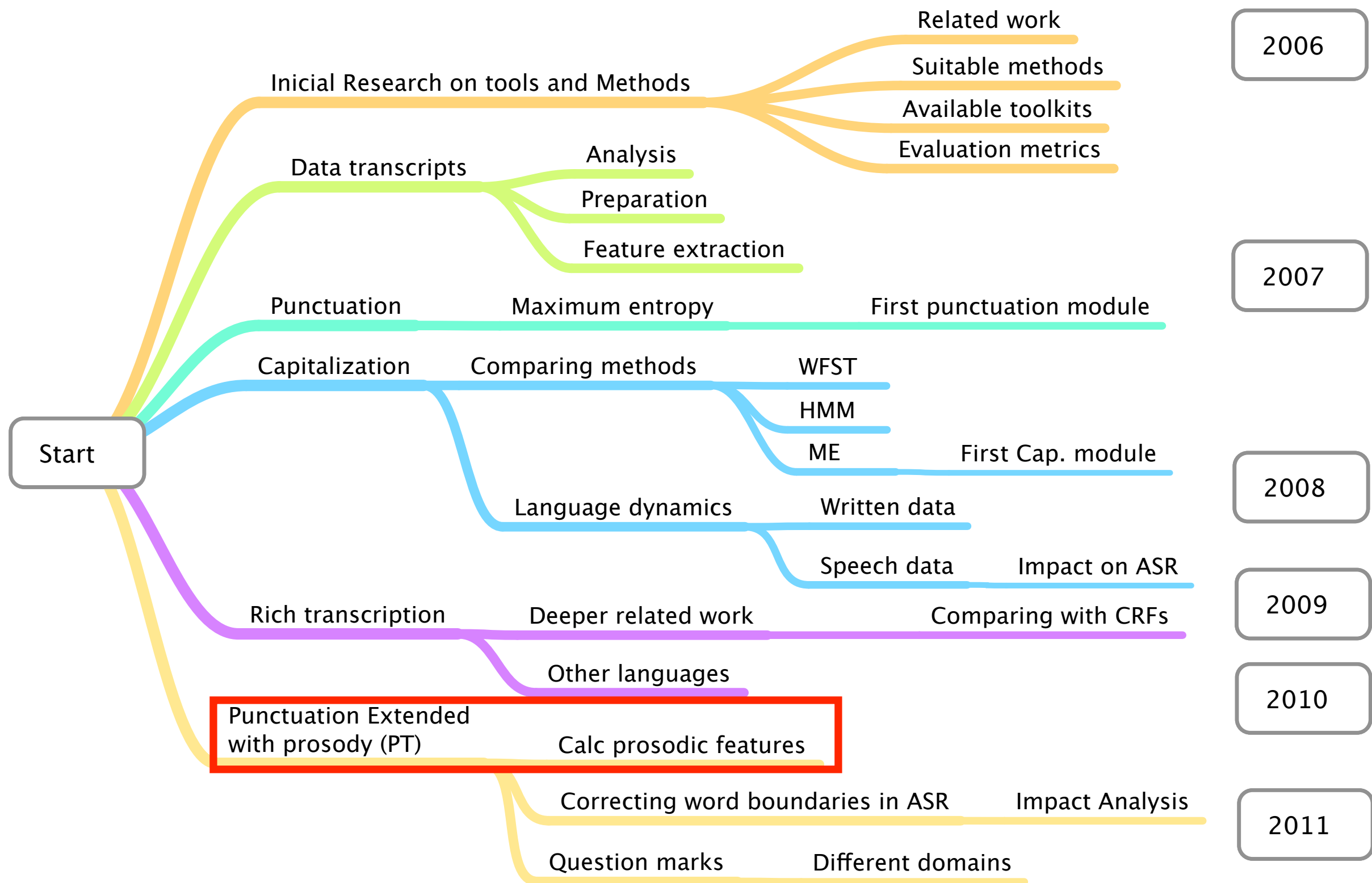


# Bilingual Experiments on Automatic Recovery of Capitalization and Punctuation of Automatic Speech Transcripts

Fernando Batista, *Member, IEEE*, Helena Moniz, Isabel Trancoso *Fellow, IEEE*, and Nuno Mamede *Member, IEEE*



# Work Overview

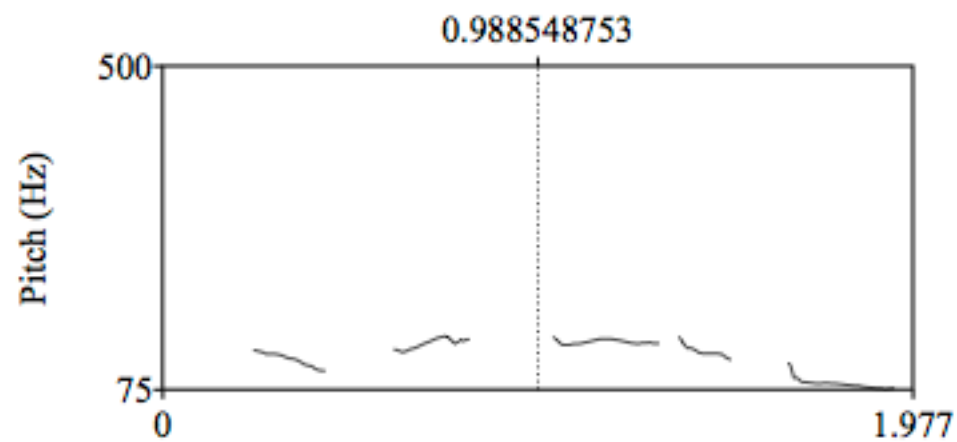
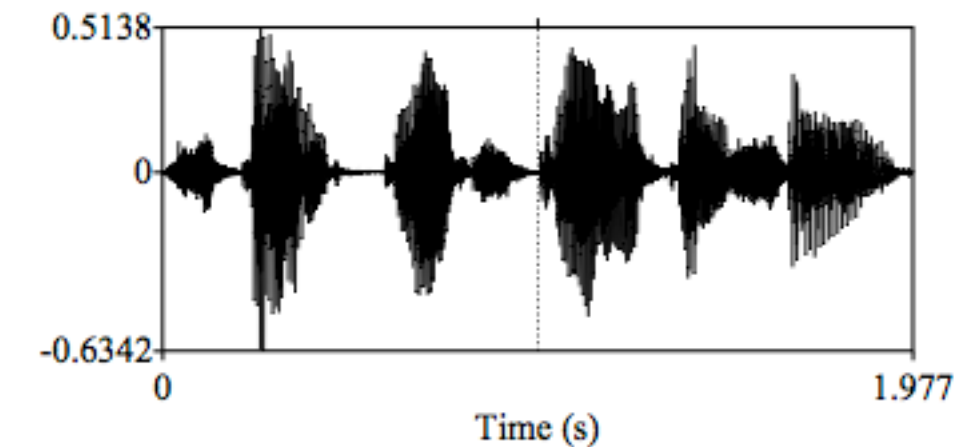


# Improving the punctuation detection

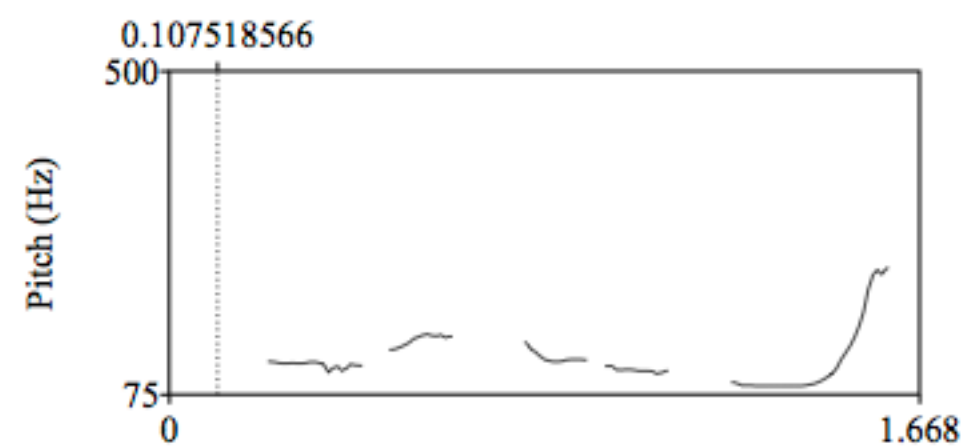
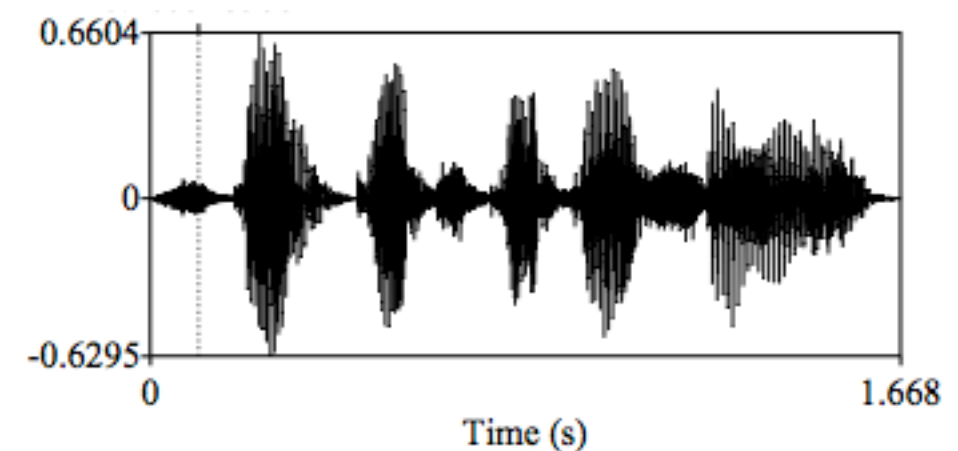
## Strategies

---

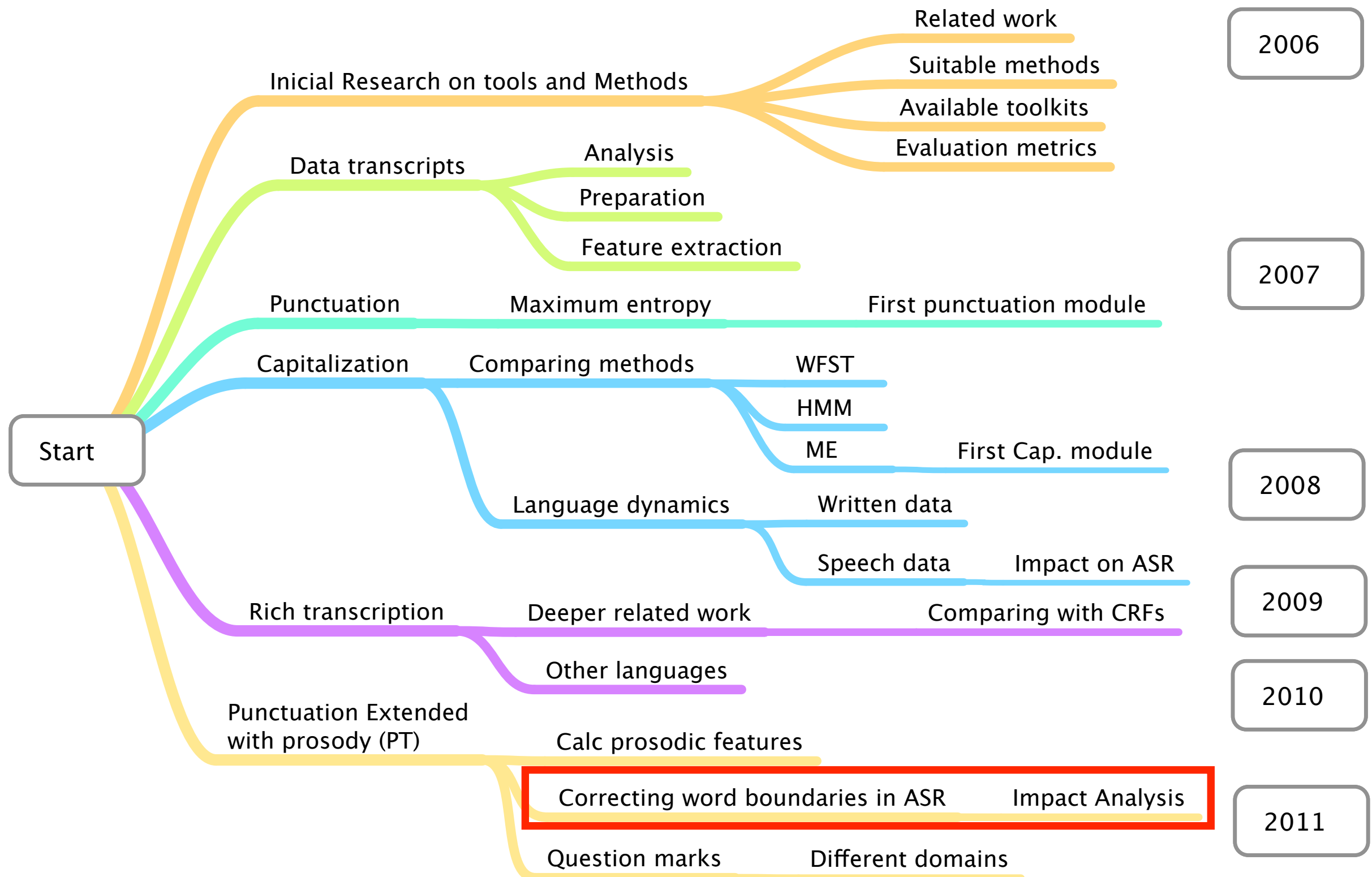
"estão todos com atenção"



"estão todos com atenção?"



# Work Overview





# Prosodically-based automatic segmentation and punctuation

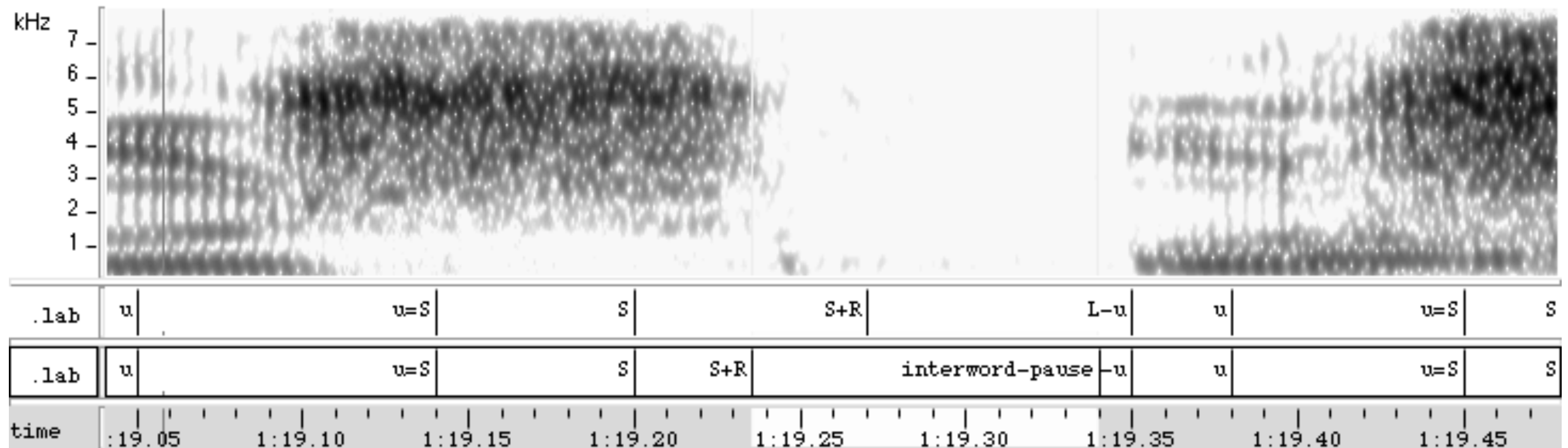
*Helena Moniz<sup>1,2</sup>, Fernando Batista<sup>2,3</sup>, Hugo Meinedo<sup>2</sup>, Alberto Abad<sup>2</sup>,  
Isabel Trancoso<sup>2</sup>, Ana Isabel Mata<sup>1</sup>, Nuno Mamede<sup>2</sup>*

<sup>1</sup>FLUL/CLUL, University of Lisbon, Portugal

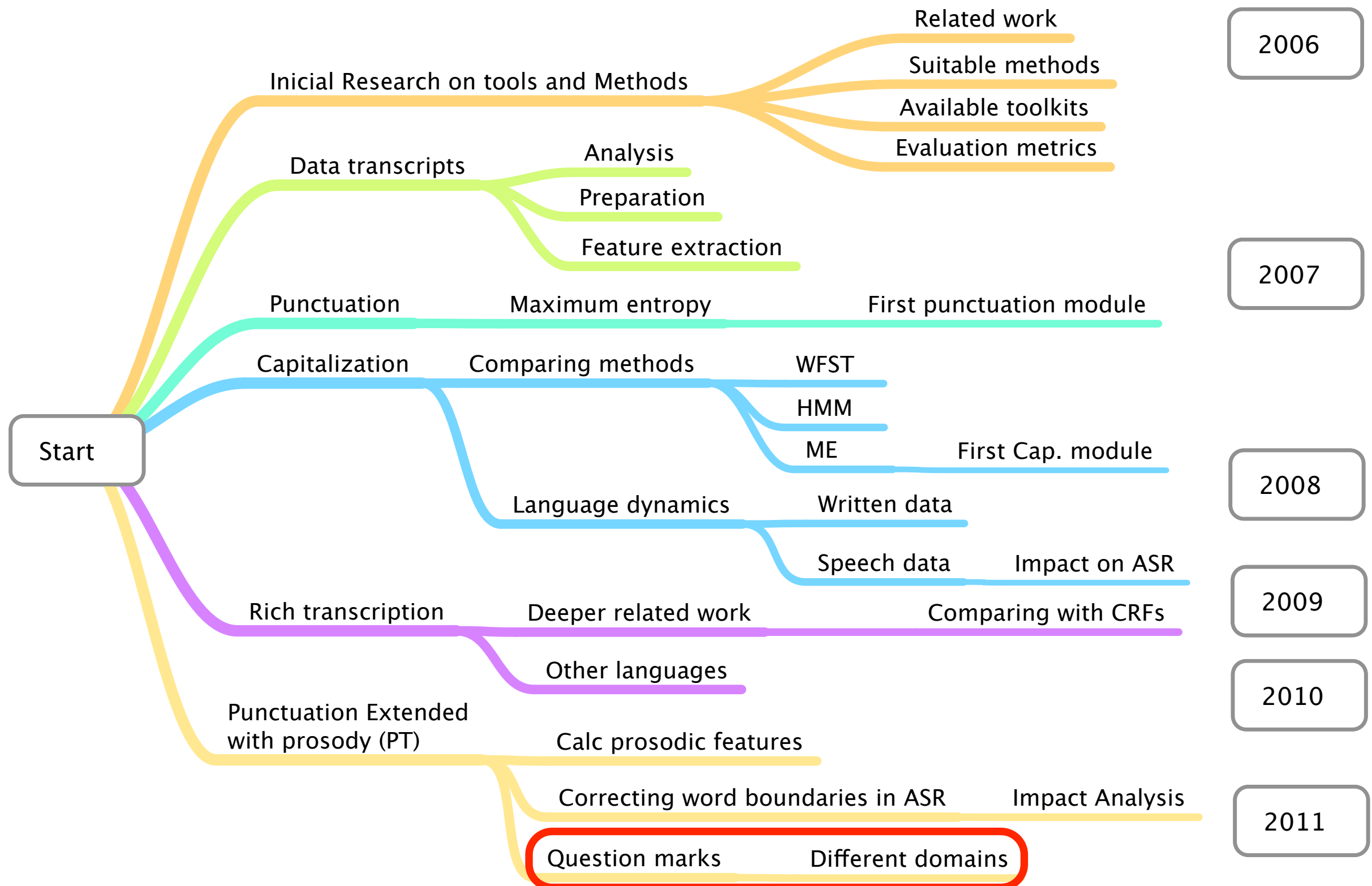
<sup>2</sup>IST / INESC-ID, Lisbon, Portugal

<sup>3</sup>ISCTE, Lisbon, Portugal

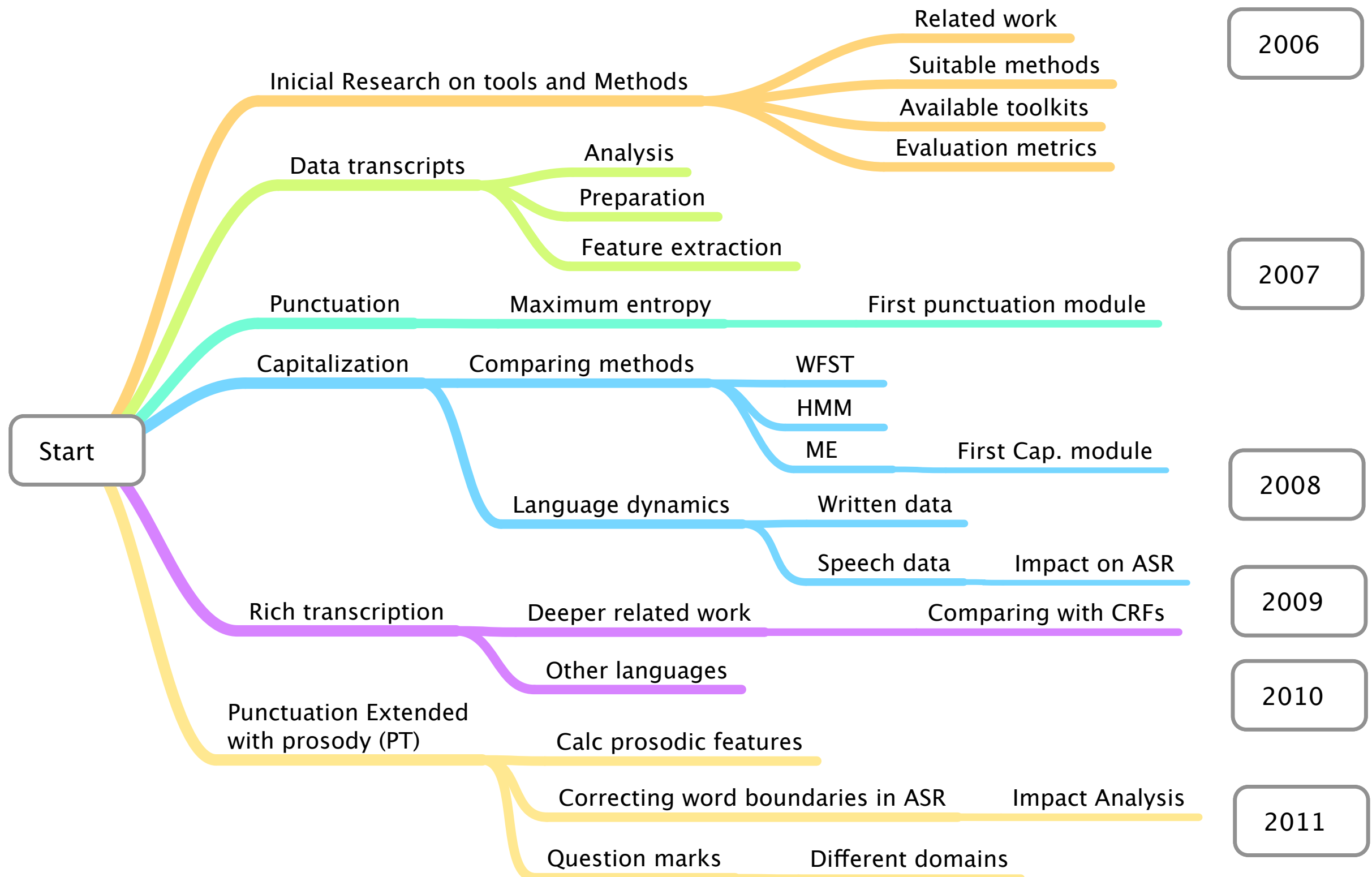
{helenam;fmmb;meinedo;alberto;isabel.trancoso;njm}@l2f.inesc-id.pt and aim@fl.ul.pt



# Work Overview



# Work Overview



# Major difficulties

---

- Multidisciplinary subjects
  - Involving: ASR, linguistics, Machine Learning, distributed systems, etc
  - Cooperation is fundamental !!!
- Too much data
  - Speech transcripts contain complex information
  - Moreover, is changing from time to time
  - At some point, it required 1 month for model training
- Things change...
  - The speech recognition system (ASR) was always evolving
  - The data changed somewhere in the middle
  - Some toolkits had bugs that were corrected in the meanwhile

# Major difficulties

## Concerning the research document

---

- Could not compare initial experiments with the final ones
  - Should I document initial experiments? say the first 3 years
  - Or ... should I simply document recent experiments?
    - comparable results, using all the good stuff, state-of-the-art results
- Solution
  - Adopt an historical perspective
    - Early work, improvements, possible variations, final results
  - Complement that with a good introduction

# Good experiences

---

- I have established a time schedule
  - weekdays, from 7h to 16h30
  - and ... I did not work on weekends and holidays ;)

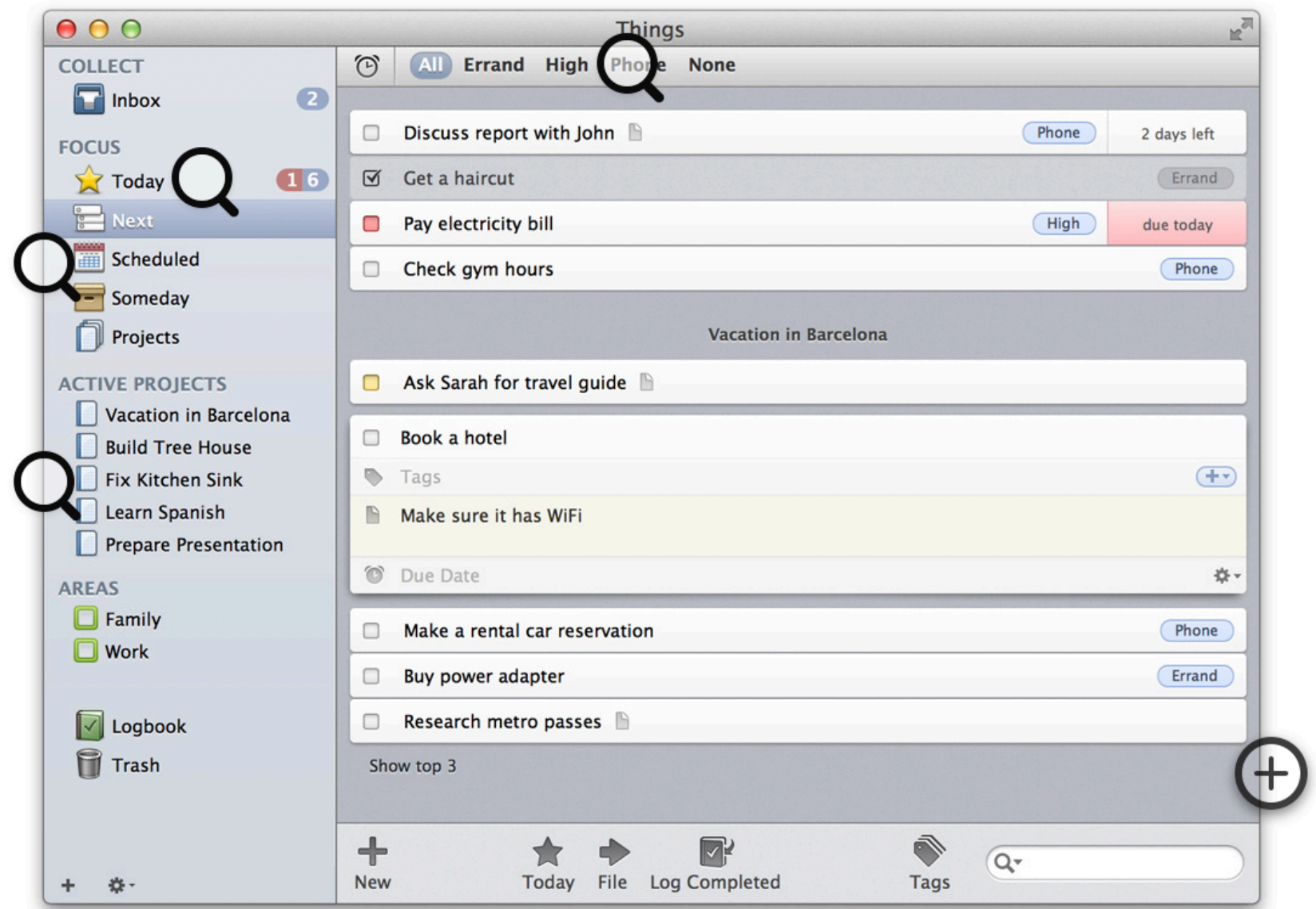


- Cooperation with other people
  - don't stay home all the time
  - opportunities, ideas, etc. arise when you deal with other people



# Some advices

- Organize your ideas and planned tasks
  - easy access to temporary notes (e.g. *sticky notes*)
  - *Things?*
  - ...



# Some advices


---

- Organize your ideas and planned tasks
  - easy access to temporary notes (e.g. *sticky notes*)
  - *Things?*
- Keep track of your research
  - bibliography manager: e.g. *Bibdesk*, *Jabref*
  - store your PDF files together
    - and annotate them with relevant information
  - personal wiki?



# Some advices ...

## Always keep track of your research



View Edit

### Research notes

#### Subjects

- [References about Social Networks](#)

#### Meetings


- [2012-01 Mini-dia do L2F](#)

#### My journals

- [Journal of Speech Sciences](#)
- IEEE Transactions on Audio, Speech and Language Processing
- Speech Communication Journal. <http://ees.elsevier.com/locate/specom>

#### References

- Machine Translation using [Moses: Tutorial, FAQ and Examples](#)
- Statistical Machine Translation [Book by Philipp Koehn](#)
- [GIT, GIT - creating a local repository](#)
- <http://pesquisa.biblioteca.iscte.pt/> - Portal Agrega
- <http://techtalks.tv/events/76/> - Videos of the 1st
- [berkeleylm](#) - A library for estimating storing large
- [Guide to diff and patch](#)
- [Punctuation and Capitalization using SRILM to](#)
- [Working with Wagon](#)
- <http://mloss.org/software/> - Machine Learning C



### Punctuation Experiments

These results were achieved with the following conditions

- bootstrapping from a previously created written corpora model (All versions bootstrap from written corpora)
- The evaluation corpora includes "eval, jeval, rtp07, and rtp08"

main conclusions

- version 2 applies recognition models for recognition data and aligned models to aligned data
  - results are a bit worse, but such models may be more interesting because are not so dependent from
  - POS info gives about 3% to 4% improvement (that may be due to the fact that FalaPosta is being used)
- Version 3 uses pseudo-syllable information and introduces a new experiment with speaker rate
  - as pseudo-silabas não causam diferenças significativas. They are slightly better for alignment (0.1%)
- The speaker rate also produces slightly worse results (0.3%).

v1-pt-alert-data.align-v1.all

tst-bn-all	bn-all	11351	2838	2541	0	80.0	81.7	80.8	38.7		11597	4949	8751	0	70.1	57.0	62.9
tst-bn-F0-F40	bn-all	6986	1437	1300	0	82.9	84.3	83.6	33.0		4987	2383	3670	0	67.7	57.6	62.1
tst-bn-F1-F41	bn-all	3249	1170	955	0	73.5	77.3	75.4	50.5		5595	2132	4266	0	72.4	56.7	63.6

v1-pt-alert-data.align-v1.lexacc

tst-bn-all	bn-all	11013	3150	2879	0	77.8	79.3	78.5	43.4		11174	4849	9174	0	69.7	54.9	61.4
tst-bn-F0-F40	bn-all	6755	1530	1531	0	81.5	81.5	81.5	36.9		4808	2348	3849	0	67.2	55.5	60.1
tst-bn-F1-F41	bn-all	3180	1354	1024	0	70.1	75.6	72.8	56.6		5373	2073	4488	0	72.2	54.5	62.1

v1-pt-alert-data.align-v1.pfeat1\_energy

tst-bn-all	bn-all	11222	3231	2670	0	77.6	80.8	79.2	42.5		11245	4863	9103	0	69.8	55.3	61.7
tst-bn-F0-F40	bn-all	6875	1556	1411	0	81.5	83.0	82.3	35.8		4849	2337	3808	0	67.5	56.0	61.1
tst-bn-F1-F41	bn-all	3242	1403	962	0	69.8	77.1	73.3	56.3		5394	2109	4467	0	71.9	54.7	62.1

v1-pt-alert-data.align-v1.pfeat1\_pitch

tst-bn-all	bn-all	11104	2559	2788	0	81.3	79.9	80.6	38.5		11566	5069	8782	0	69.5	56.8	62.5
tst-bn-F0-F40	bn-all	6833	1298	1453	0	84.0	82.5	83.2	33.2		4984	2451	3673	0	67.0	57.6	61.1
tst-bn-F1-F41	bn-all	3187	1064	1017	0	75.0	75.8	75.4	49.5		5566	2158	4295	0	72.1	56.4	63.1

v1-pt-alert-data.align-v1.pfeat1

tst-bn-all	bn-all	11522	3264	2370	0	77.9	82.9	80.4	40.6		11318	4744	9030	0	70.5	55.6	62.2
tst-bn-F0-F40	bn-all	7098	1650	1188	0	81.1	85.7	83.3	34.3		4845	2254	3812	0	68.2	56.0	61.1
tst-bn-F1-F41	bn-all	3297	1354	907	0	70.9	78.4	74.5	53.8		5467	2057	4394	0	72.7	55.4	62.9

# Questions ?

---

Obrigado!