<insert title>

Projet d'approfondissement (PA)

Lucy Linder
12.03.2018

Mi name isch …

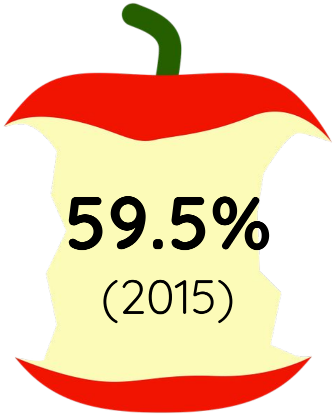
DAPLAB
Data Analysis and Processing Lab

BIG DATA

GDG
Fribourg

Grüezi… Chuchichäschtli !

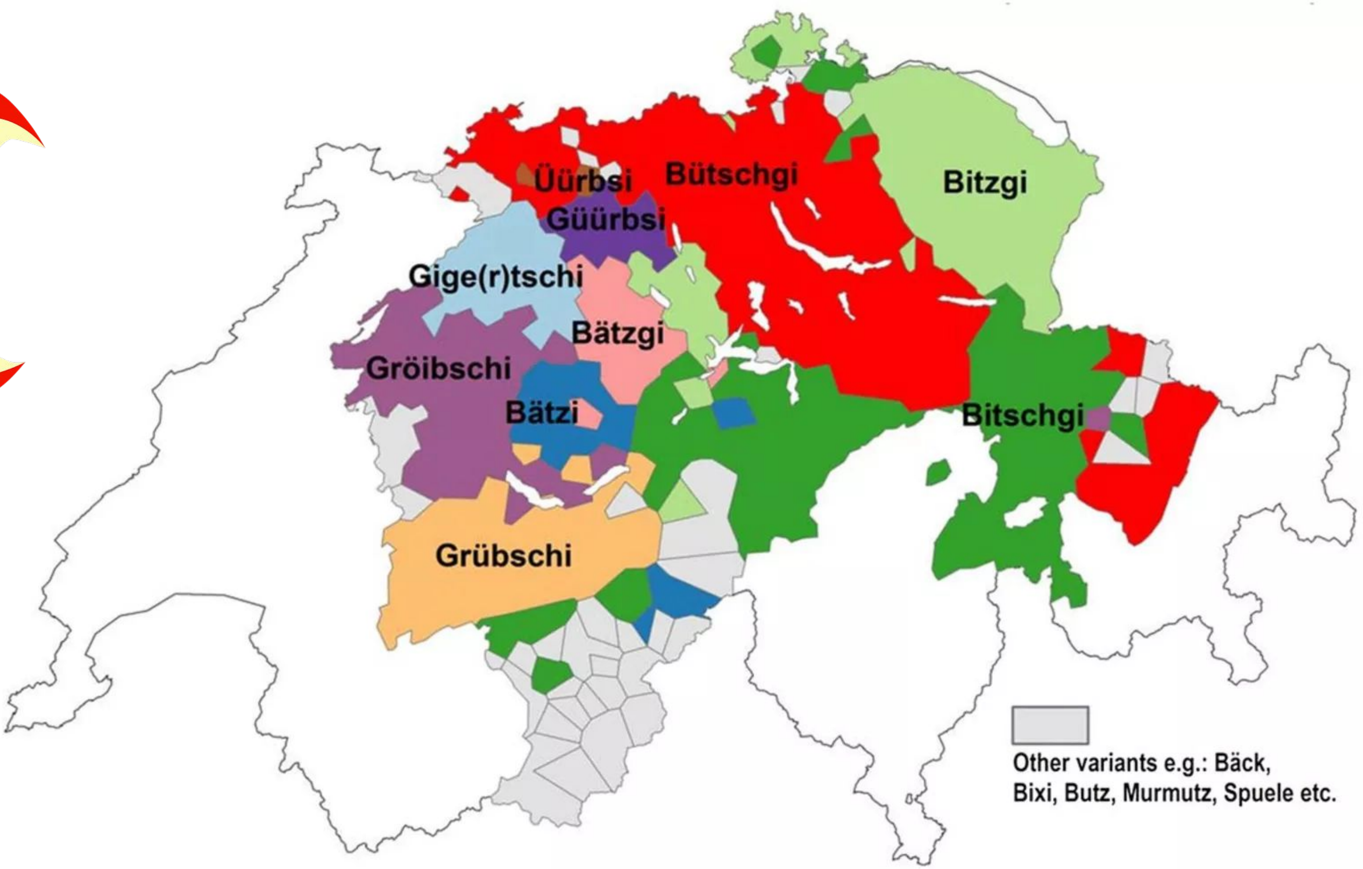
Swiss German Dialects… *This* important ? YES.



59.5%

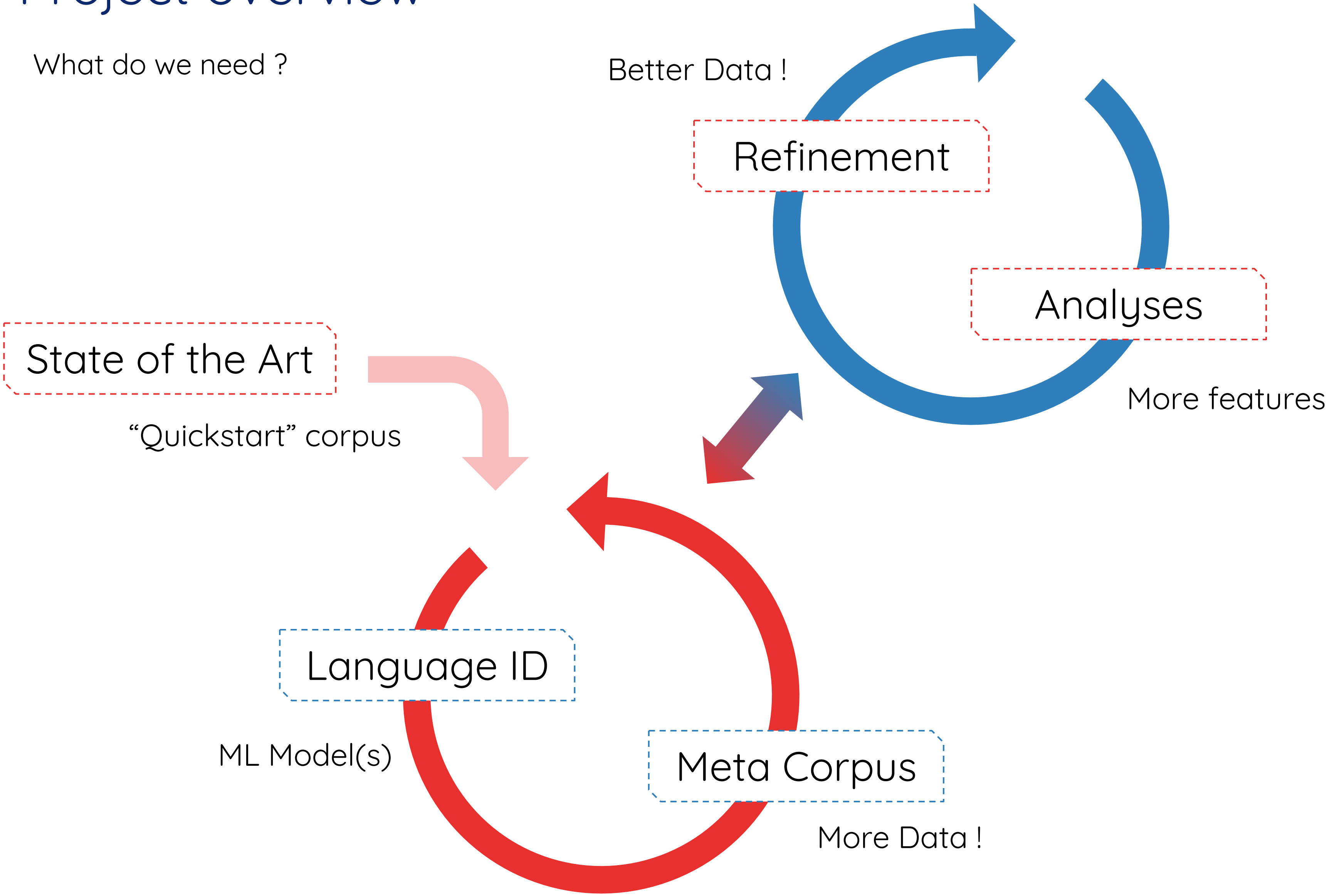
(2015)

0.066% worldwide



Project overview

What do we need ?



Better Data !

Refinement

Analyses

More features

State of the Art

“Quickstart” corpus

Language ID

ML Model(s)

Meta Corpus

More Data !

State of the Art

Existing corpora

NOAH's Corpus of Swiss German Dialects

→ 7’000 sentences, 115’000 tokens

ArchiMob corpus

→ 528’381 tokens

Sms4science

→ 10’800 sms in SG, 288’400 tokens

SB-CH corpus

→ ?

nearly

1 million

tokens

And many more to scrape

State of the Art

NOAH's Corpus of Swiss German Dialects

→ 5 sources

→ 2 publications, 2014-2015

Alemannic Wikipedia	22'136
Swatch Annual Report 2012	33'023
Novels from Viktor Schobinger	12'858
Newspaper articles (Blick)	11'256
Blogs	34'404
	113'677

```
<document dialect="various" title="?">
 <article n="0" dialect="various"
          title="Bettwanze und Vire im Gepäck">
  <s n="0-0">
   <w n="0-0-0" pos="NN">Bettwanze</w>
   <w n="0-0-1" pos="KON">und</w>
   <w n="0-0-2" pos="NN">Vire</w>
   <w n="0-0-3" pos="APPRART">im</w>
   <w n="0-0-4" pos="NN">Gepäck</w>
   <w n="0-0-5" pos="NN">Vorsorg</w>
   …
```

→ POS-tagging: Stuttgart-Tubingen-TagSet (STTS) + PTKINF

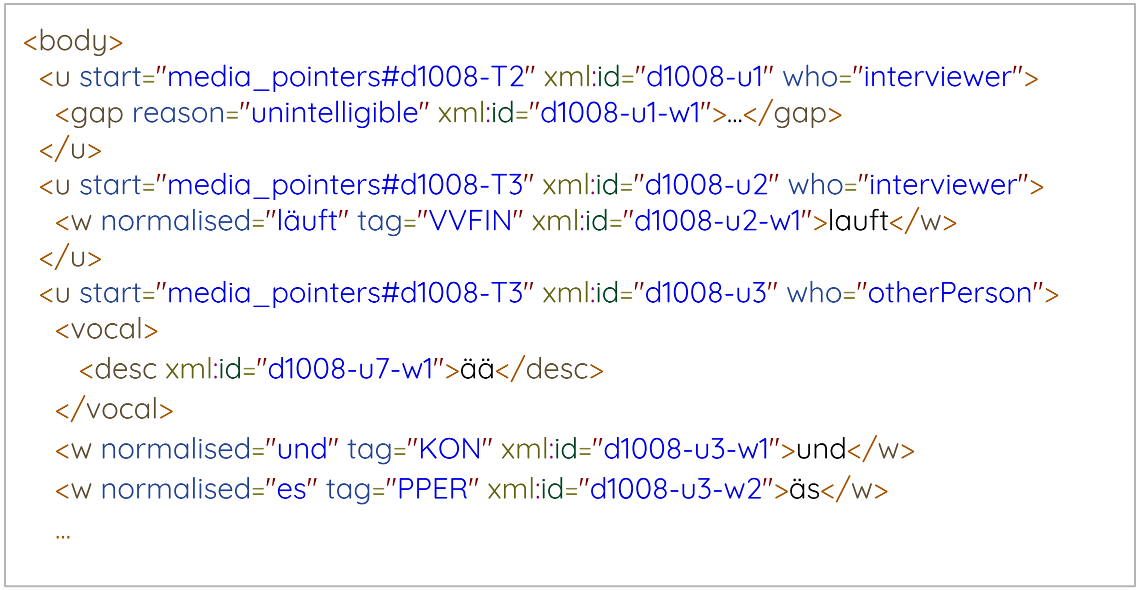
→ *BTagger* (demo)

statistical Part-of-Speech tagger
accuracy: 90.62%

State of the Art

ArchiMob corpus

→ long samples of transcribed text, speech-to-text alignment



```
<body>
 <u start="media_pointers#d1008-T2" xml:id="d1008-u1" who="interviewer">
  <gap reason="unintelligible" xml:id="d1008-u1-w1">…</gap>
 </u>
 <u start="media_pointers#d1008-T3" xml:id="d1008-u2" who="interviewer">
  <w normalised="läuft" tag="VVFIN" xml:id="d1008-u2-w1">lauft</w>
 </u>
 <u start="media_pointers#d1008-T3" xml:id="d1008-u3" who="otherPerson">
  <vocal>
    <desc xml:id="d1008-u7-w1">ää</desc>
  </vocal>
  <w normalised="und" tag="KON" xml:id="d1008-u3-w1">und</w>
  <w normalised="es" tag="PPER" xml:id="d1008-u3-w2">äs</w>
  …
```

→ automatic normalisation (accuracy: 90.46%)

→ automatic Part-of-Speech tagger (accuracy: 90.09%)

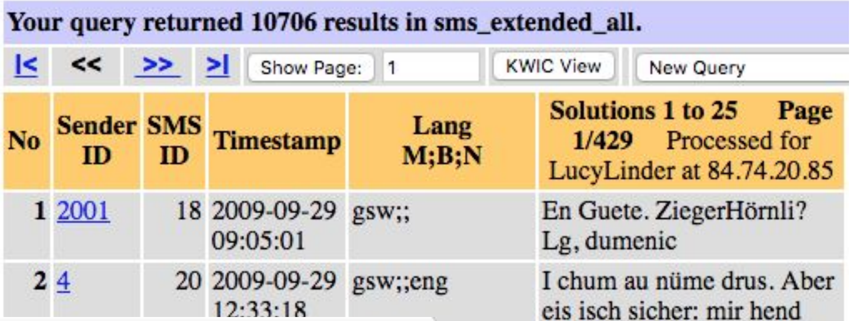
State of the Art

Sms4science

→ 2 corpora:

“raw” vs normalised

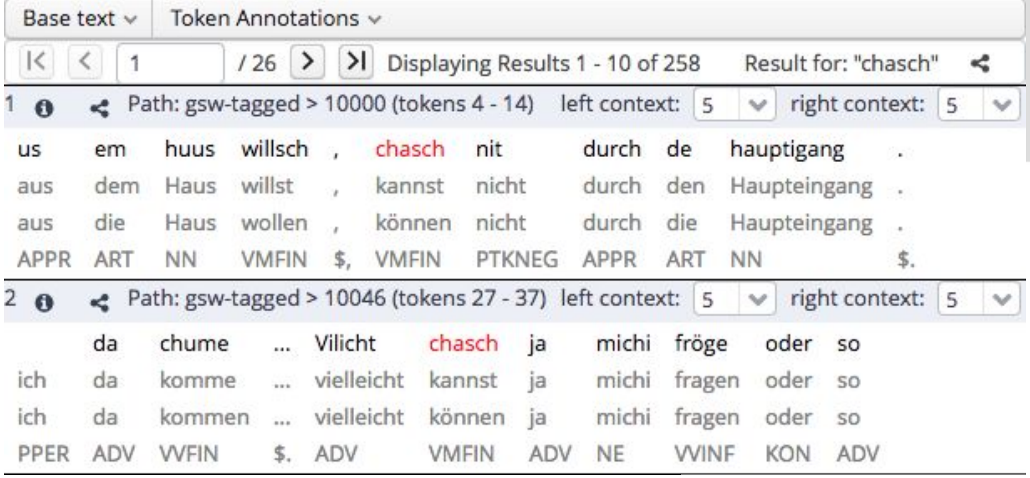
→ needs login



Your query returned 10706 results in sms_extended_all.

No	Sender ID	SMS ID	Timestamp	Lang M;B;N	Solutions 1 to 25 Page 1/429 Processed for LucyLinder at 84.74.20.85
1	2001	18	2009-09-29 09:05:01	gsw;;	En Guete. ZiegerHörnli? Lg, dumenic
2	4	20	2009-09-29 12:33:18	gsw;;eng	I chum au nüme drus. Aber eis isch sicher: mir hend hüt vom 2-viertelvor 4i vorlesig im kol-f-NNN, ob [LastName] oder dä ander. Morn hemer ka ta, stoht:"no seßion"

Sms4science navigator



Displaying Results 1 - 10 of 258 Result for: "chasch"

1 Path: gsw-tagged > 10000 (tokens 4 - 14) left context: 5 right context: 5

us	em	huus	willsch	,	chasch	nit	durch	de	hauptigang	.
aus	dem	Haus	willst	,	kannst	nicht	durch	den	Haupteingang	.
aus	die	Haus	wollen	,	können	nicht	durch	die	Haupteingang	.
APPR	ART	NN	VMFIN	$,	VMFIN	PTKNEG	APPR	ART	NN	$.

2 Path: gsw-tagged > 10046 (tokens 27 - 37) left context: 5 right context: 5

	da	chume	...	Vilicht	chasch	ja	michi	fröge	oder	so
ich	da	komme	...	vielleicht	kannst	ja	michi	fragen	oder	so
ich	da	kommen	...	vielleicht	können	ja	michi	fragen	oder	so
PPER	ADV	VVFIN	$.	ADV	VMFIN	ADV	NE	VVINF	KON	ADV

ANNIS interface

→ normalisation and PoS tagging

Many steps by hand

STTS tagset + PTKINF, TreeTagger with tailor-made parameter

FNSNF Swiss National Science Foundation

Language ID

Using machine learning

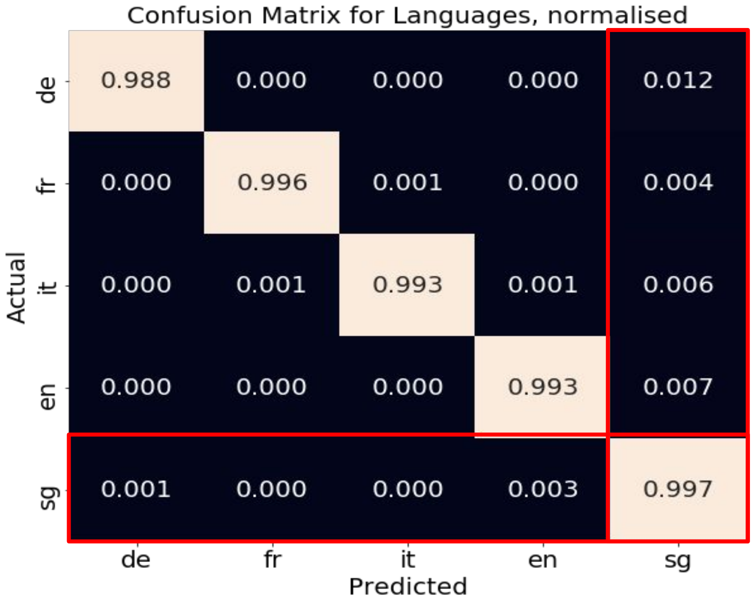
Model and data

→ Character level

ngrams of variable size: 1-6

→ 7’000 sentences per langage

From europarl + sms4science



	de	fr	it	en	sg
de	0.988	0.000	0.000	0.000	0.012
fr	0.000	0.996	0.001	0.000	0.004
it	0.000	0.001	0.993	0.001	0.006
en	0.000	0.000	0.000	0.993	0.007
sg	0.001	0.000	0.000	0.003	0.997

Results

	precision	recall	f1-score	support
sg	**0.97**	1.00	**0.98**	1451
fr	1.00	1.00	1.00	1370
it	1.00	0.99	1.00	1366
en	1.00	0.99	0.99	1421
de	1.00	0.99	0.99	1392

Sample of misclassification

real	pred	sentence
fr	sg	Rapport Schroedter (A5-0108/1999)
it	sg	.
de	sg	Kultur 2000
sg	en	Go out the way, i’m coming round the corner!

Perspectives

Data sources

Web: Twitter, blogs, …

National library: books, articles, …

Language ID

N-grams variations

CNNs

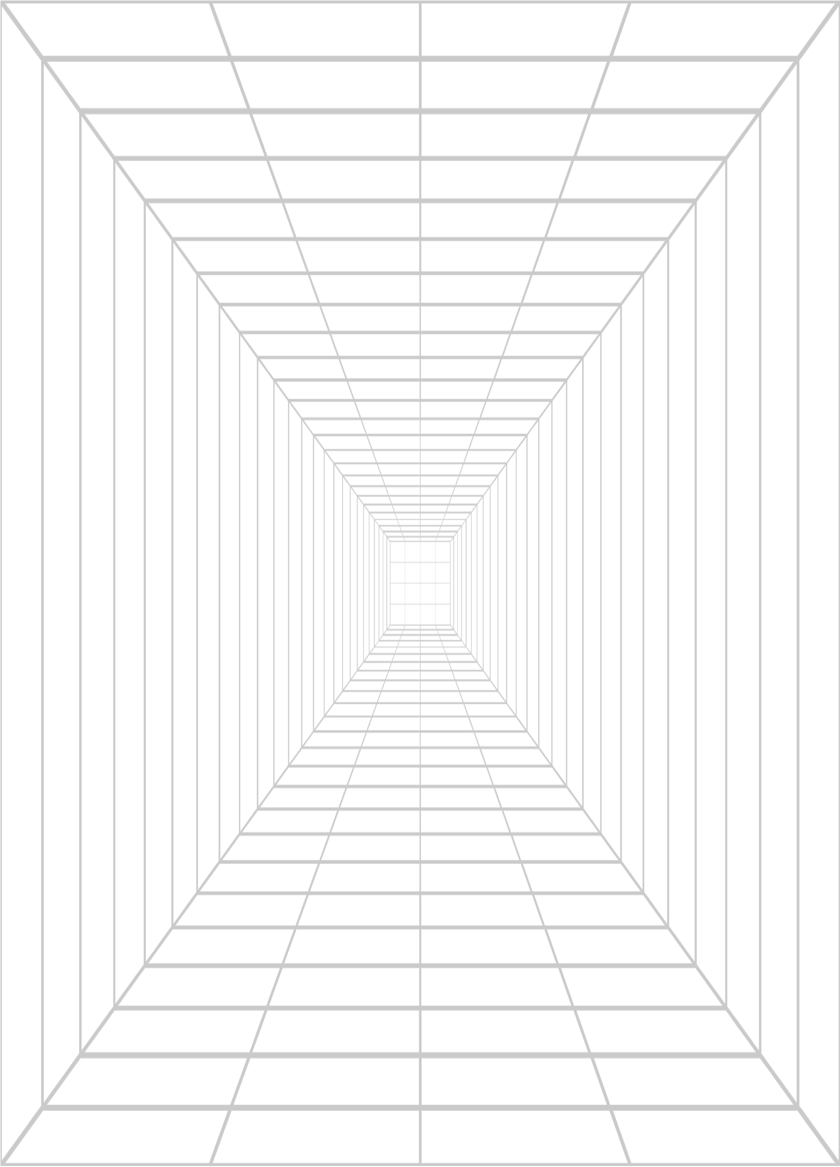
… Detecting dialects ?

Other analyses

Sentiment analysis

Automatic normalisation

…



Project outcomes

Summary

- State of the Art

- Meta/bigger Corpus

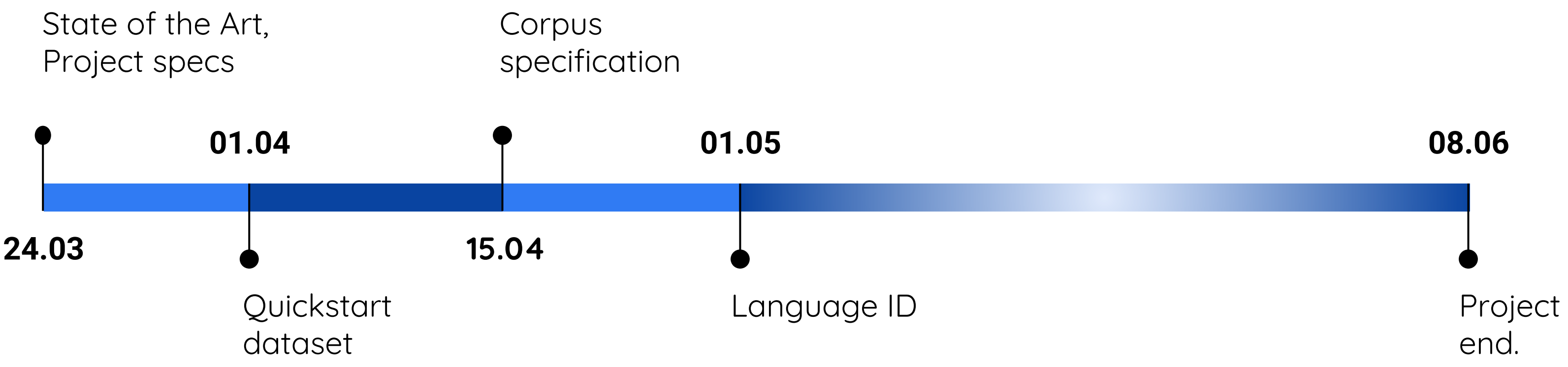
- Language identifier(s)

- Other analyses (nice-to-have)

Calendar

Agile methodology

Major milestones



State of the Art, Project specs

24.03

01.04

Quickstart dataset

Corpus specification

15.04

01.05

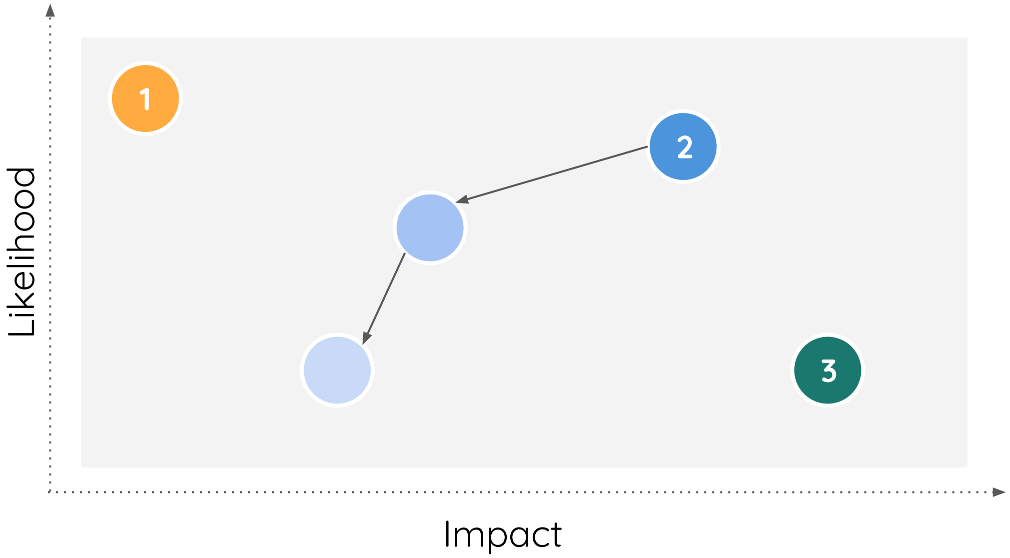
Language ID

08.06

Project end.

One last thing

Risks



1 I don’t speak SG

2 Unclear goals

Agile

Discussions with you

3 Your needs already satisfied

Merci vilmal

What are **YOUR** needs ?

Priorities and outcomes ?

Corpus base unit ?

...