Speech is Golden

- on ASR at the service of the Danish public sector

Peter Juel Henrichsen

DanCAST - Danish Center for Applied Speech Technology
Copenhagen Business School
University of Copenhagen

pjh.ibc@cbs.dk

	🇸🇮	🇩🇰
L1 speakers	2.2 mio	5.5 mio
EU working language	yes	yes
Mix of municipalities	5-6 city / many small	4 city / 94 small

L1 speakers	2.2 mio	5.5 mio
EU working language	yes	yes
Mix of municipalities	5-6 city / many small	4 city / 94 small

	Slovene	**Danish**
Inflected language	>>English	>English
Compounding	>English	>English
Rich in vowel qualities	>English	>>English

This talk

1. Why ASR in the municipalities?
2. ASR - the technology
3. Trough of disillusionment
4. The new alliance

Why ASR in the Danish municipalities?

The local authorities smelled a business case...

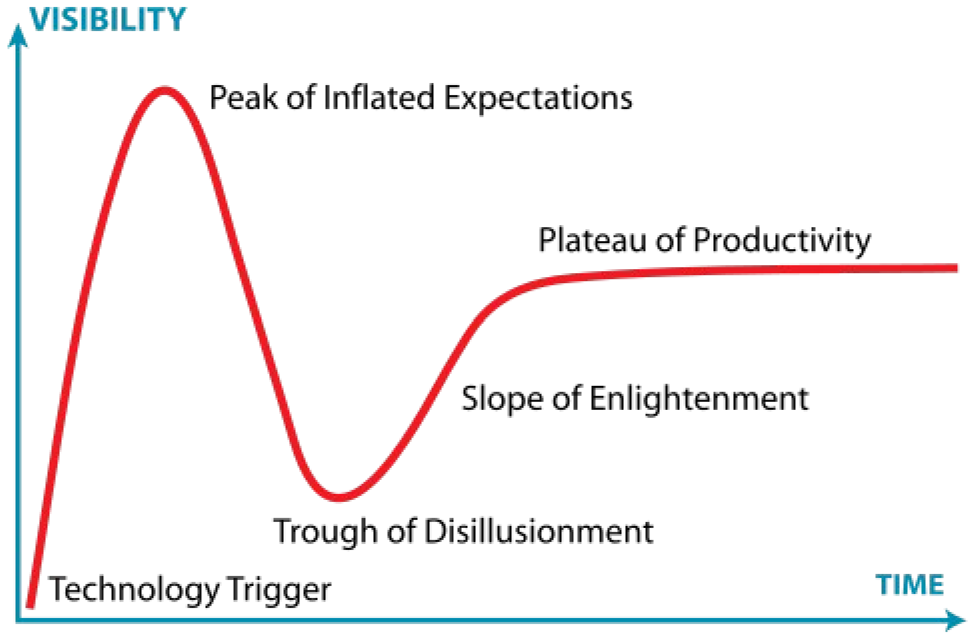
The slick salesman:

1) economy! - the medical case
2) speech aid for the challenged
3) 'welfare tech' - assisting in difficult working situations

The technology looked extremely user friendly and mature

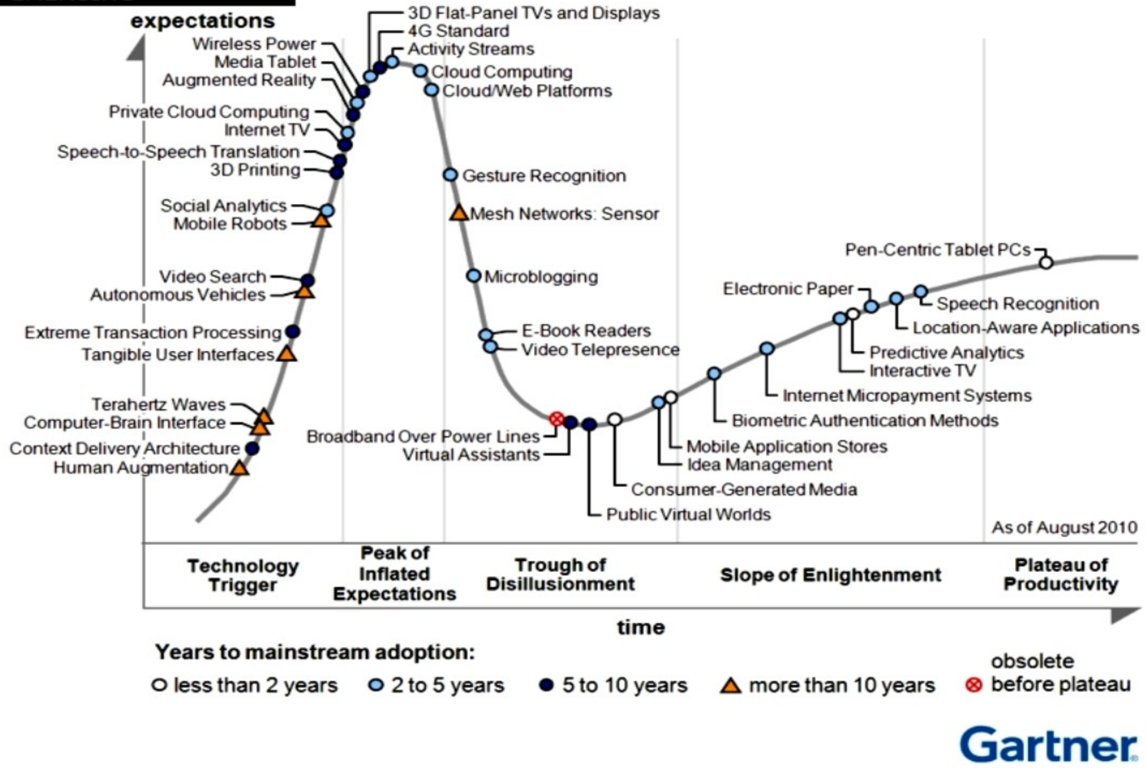
[ad]

On top of that, the *real* world had embraced ASR already



Gartner hype curve

2010 EMERGING



tazti speech recognition software
Now on Sale
Sale: ~~$80.00~~ $39.99
BUY NOW

All NEW
Dragon Dictate
for Mac, v4
Dragon Dictate
for Mac
NUANCE

However, it did not go so smoothly!

A quick intro to ASR

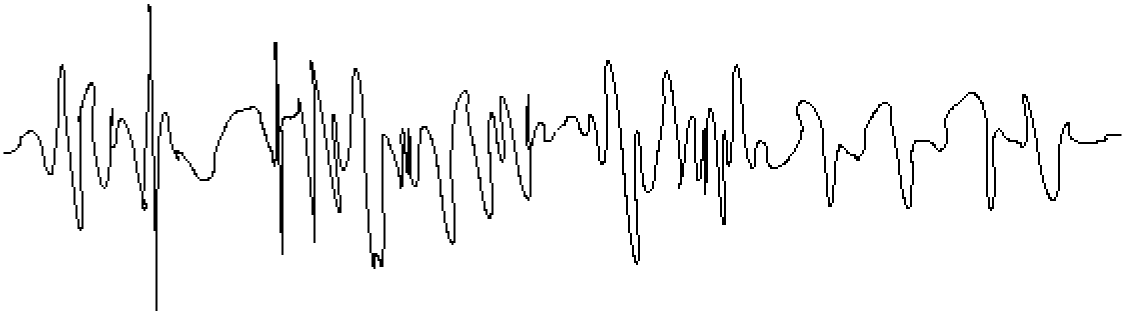
ASR - the technology

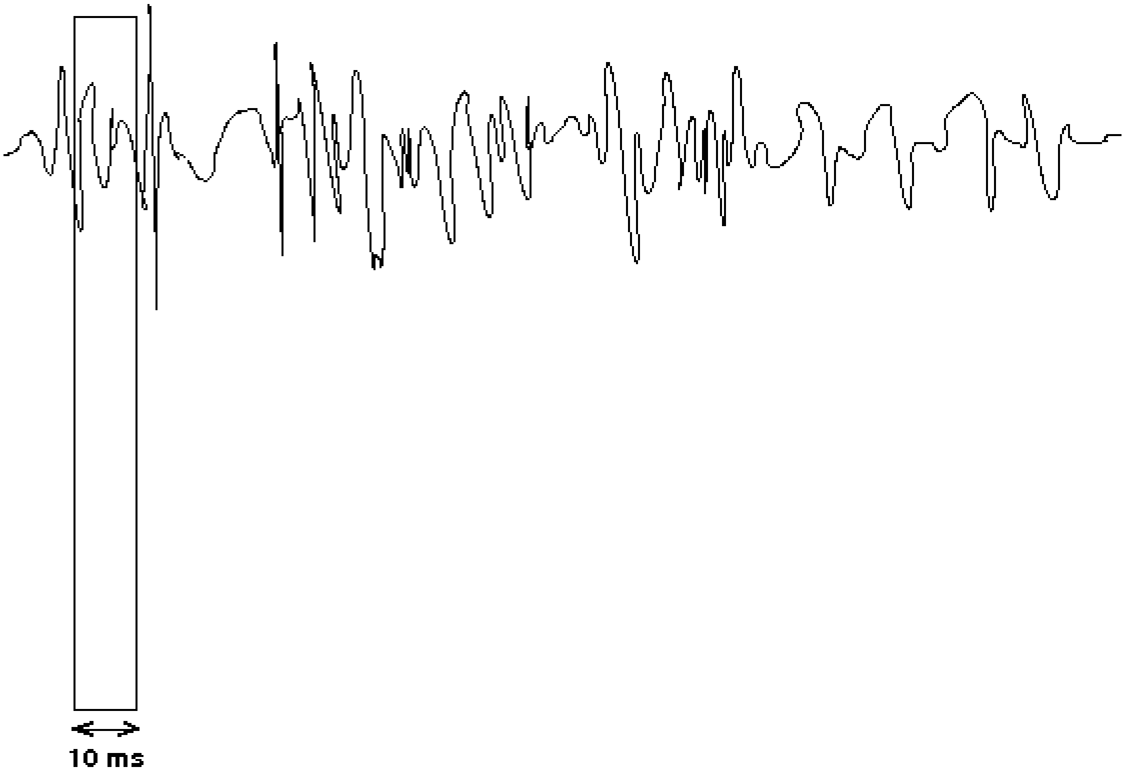
Three central components

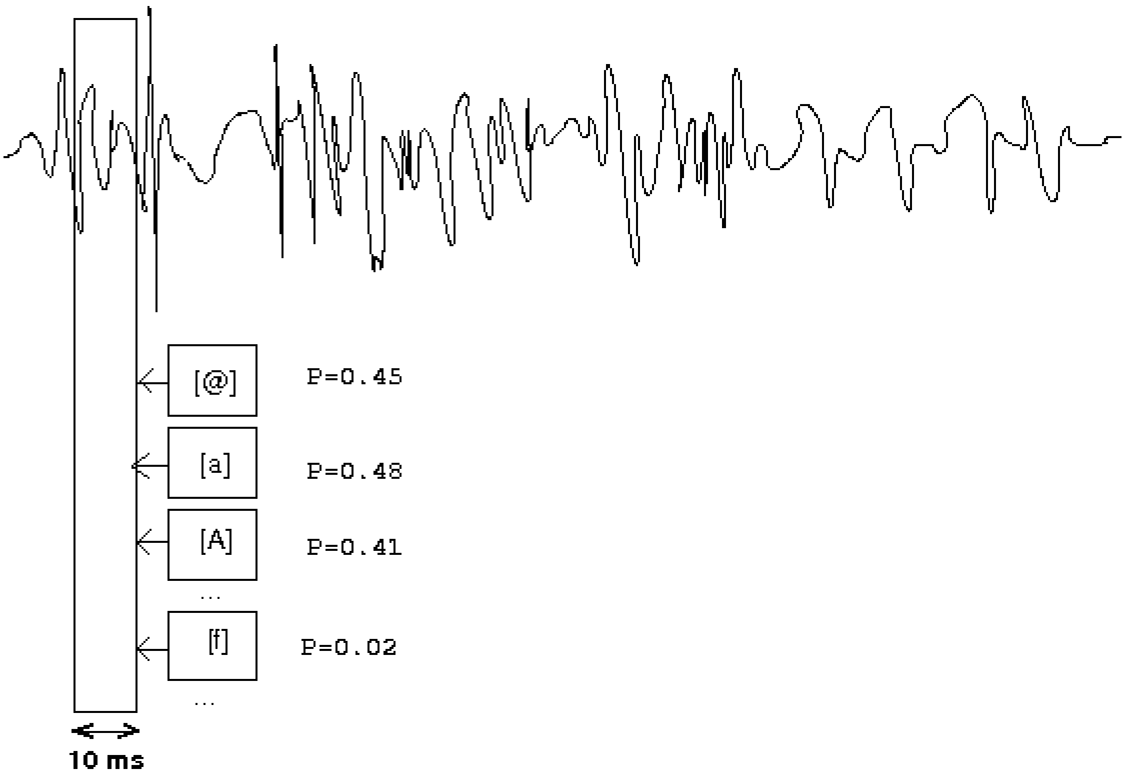
- Acoustic model (**AM**)
- Language model (**LM**)
- Search engine referring to **AM** og **LM**

Acoustic model

A set of recognizers, one for each language sound ('phone')




10 ms


[@]
P=0.45
[a]
P=0.48
[A]
P=0.41
...
[f]
P=0.02
...
10 ms

Language model

- Unigrams
- Bigrams
- Trigrams
- (n-grams)

all frequency annotated (NB! corpus-driven)

UNIGRAMS:
the > is > in > cat > oven

BIGRAMS:
the cat > the fat >>> *the that >>> **the sat

Language model

- Unigrams
- Bigrams
- Trigrams
- (n-grams)

all frequency annotated (NB! corpus-driven)

UNIGRAMS:
the > is > in > cat > oven

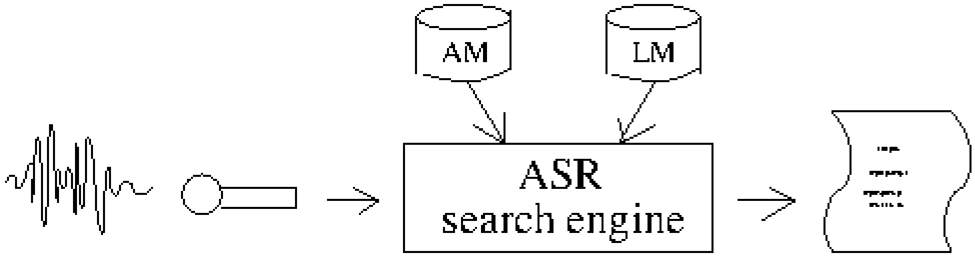
BIGRAMS:

the cat	About 89.400.000 results	(by Google)
the fat	About 40.100.000 results	
the that	About 10.300.000 results	
the sat	About 7.020.000 results	

TRIGRAMS:
is in the >> in the oven >> ...

N-GRAMS (domain specific mwe.s)
the cat is on the mat
the cat is in the oven

ASR: a search engine over AM and LM



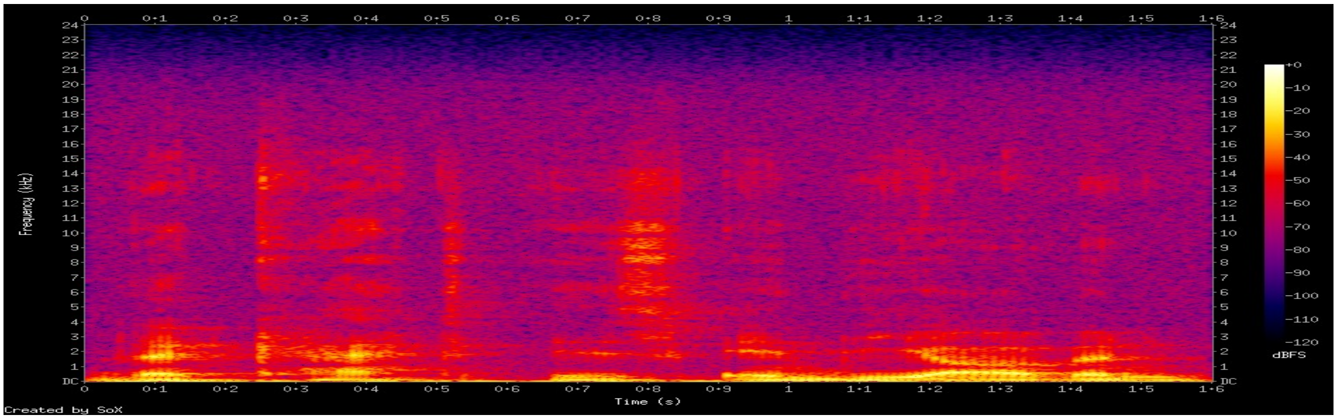
How are the AM and the LM developed?

Based on annotated corpus data

(tokenized, transcribed, tagged, time-coded)

Example: the speech corpus

[rec]



↑ the ↑ cat ↑ is ↑ in ↑ the ↔ ↑ oven

the	cat	is	in	the	oven
DH AH0	K AE1 T	IH1 Z	IH0 N	DH IH1	AH1-V-AH0-N
0 140	170 210 280	350 480	680 810	1010 1260	1310 **????**

(phonetic script: cmu dict)

[play]

Training data

Acoustic model materials	**size (order of mag.)**
phonetic dictionary	100,000 lemmas
speech recordings (multi-speaker)	100 hours
speech recordings (focus users)	1 hour each

Language model materials	**size (order of mag.)**
text corpus (general)	100M words
text corpus (specific for professional area)	100k words
non-linguistic tokens (forms, symbols, ...)	100 documents

back on track...

Status as of 2011

ASR contracts by 2011

Vendors:	**4**
KMD IBM Denmark PDC Dictus Max Manus	
Technological suppliers:	**1**
Investors (municipalities):	**22**
Positive business-cases:	

ASR contracts by 2011

Vendors:	**4**
KMD IBM Denmark PDC Dictus Max Manus	
Technological suppliers:	**1**
Investors (municipalities):	**22**
Positive business-cases:	**0**

Did Gartner lie?

Maybe not,

but several things conspired against the municipalities:

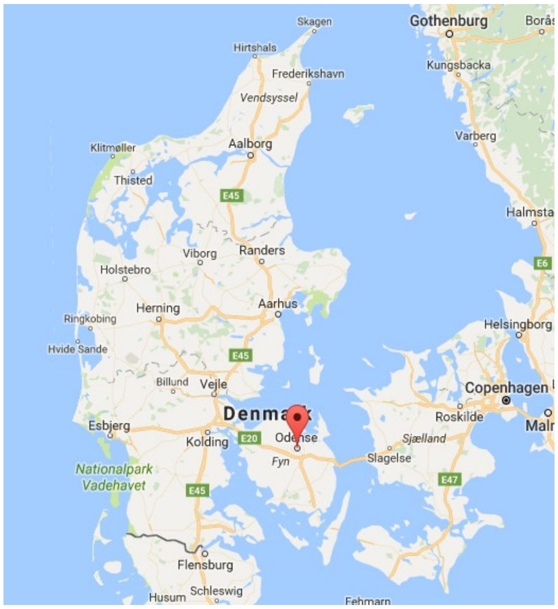
- Danish has difficult words: long, inflected, ...
- Danish has difficult vowels: lenitions, reductions, assimilations, ...

- MONOPOLY

The case of Odense

(they did everything right)

Baltic Sea
Sea
Denmark
Hamburg
Berlin
Poland
Wars
msterdam
Netherlands
ussels
Cologne
Germany
Belgium
Frankfurt
Luxembourg
Prague
Czech Republic
Munich
Vienna
Slovakia
Budapest
Austria
Hungary
Switzerland
Slovenia
Zagreb
Milan
Croatia
Belgrade
Београд
Bosnia and Herzegovina
Serbi
Sarajevo
Monaco
Italy



Odense features

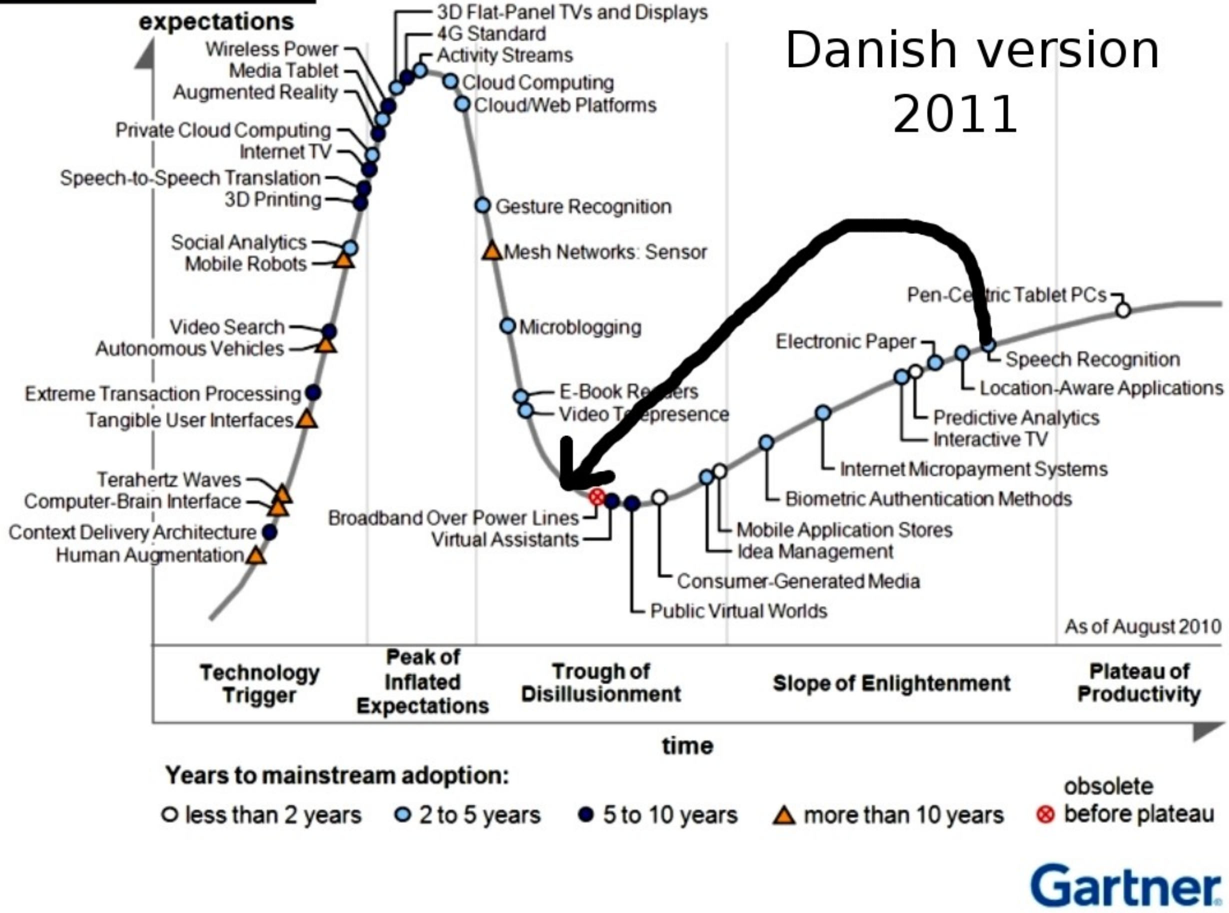
- Three year budget
- Dedicated project leader
- Training programme for employees
- 900+ participants = North-Europe's biggest

Odense did everything right - and failed

- Inflated expectations - 'savings' already entered in next-year budget!
- No clear HR policy (few employees liked ASR, most gave up)
- Vendors soon vanished: Poor service after contract was signed
- Extremely slow updates (waiting 24 months for new context files)

At project end, ~150 active users (<20%)

2010 EMERGING



Danish version
2011

The new alliance

The new alliance

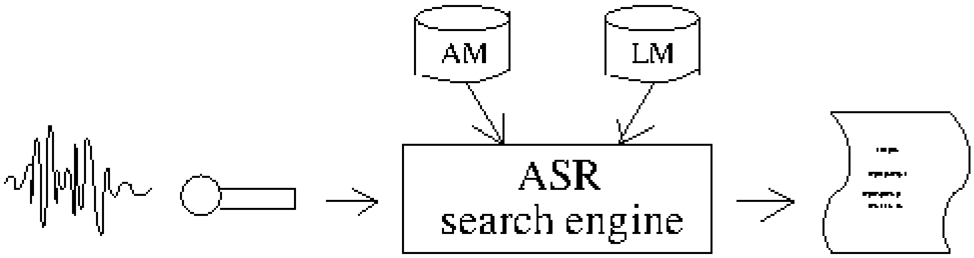
Steering committee

- OS2 (50+ Danish municipalities, 100% flat organization)
- DanCAST (Copenhagen Business School)

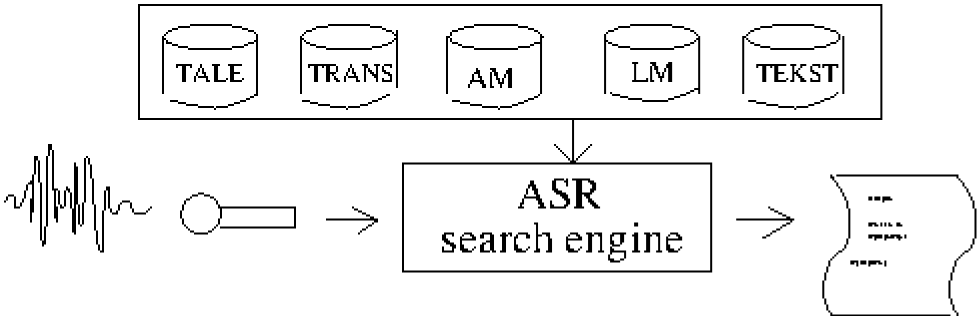
Advisory board

- KOMBIT (independent advisor)
- CST (Copenhagen University)
- Danish Parliament (Folketinget)
- Danish National Broadcast (Danmarks Radio)

First action point: Recycling of resources



Opaque module structure



Similar experiences everywhere

- all keep paying for the same resources (e.g. phonetic lexicon)
- loss of ownership to own data (e.g. annotated speech files)
- licence lock-in (change product = begin from scratch)
- no knowledge transfer (data exchange barred)

Recycling corpus materials

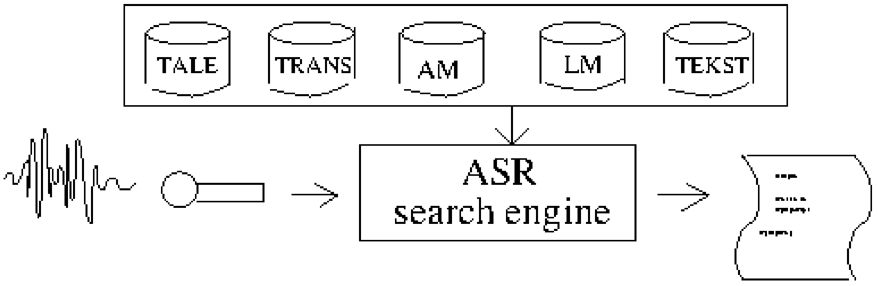
Acoustic model materials	**size (order of mag.)**	**recycled?**
phonetic dictionary	100,000 lemmas	yes
speech recordings (multi-speaker)	100 hours	yes
speech recordings (specific users)	1 hour each	no

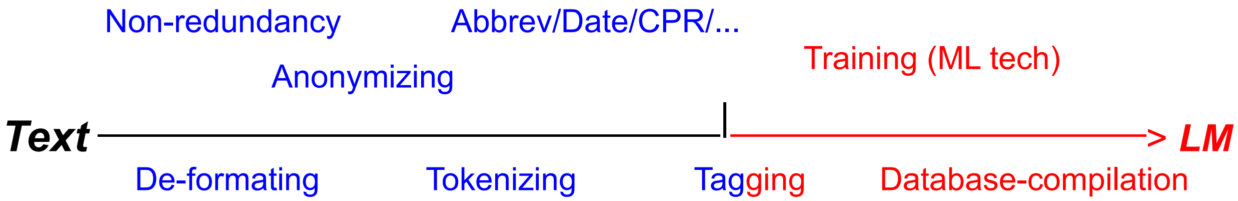
Language model materials	**size (order of mag.)**	**recycled?**
text corpus (general)	100,000,000 words	yes
text corpus (specific)	100,000 words	no
non-linguistic tokens (forms, symbols, ...)	100 documents	partly

Optimal recycling of training data

Recyclable : Project specific = **50 : 1**

Drawing the line between mine and yours





Non-redundancy

Abbrev/Date/CPR/...

Anonymizing

Training (ML tech)

Text

> ***LM***

De-formating

Tokenizing

Tagging

Database-compilation

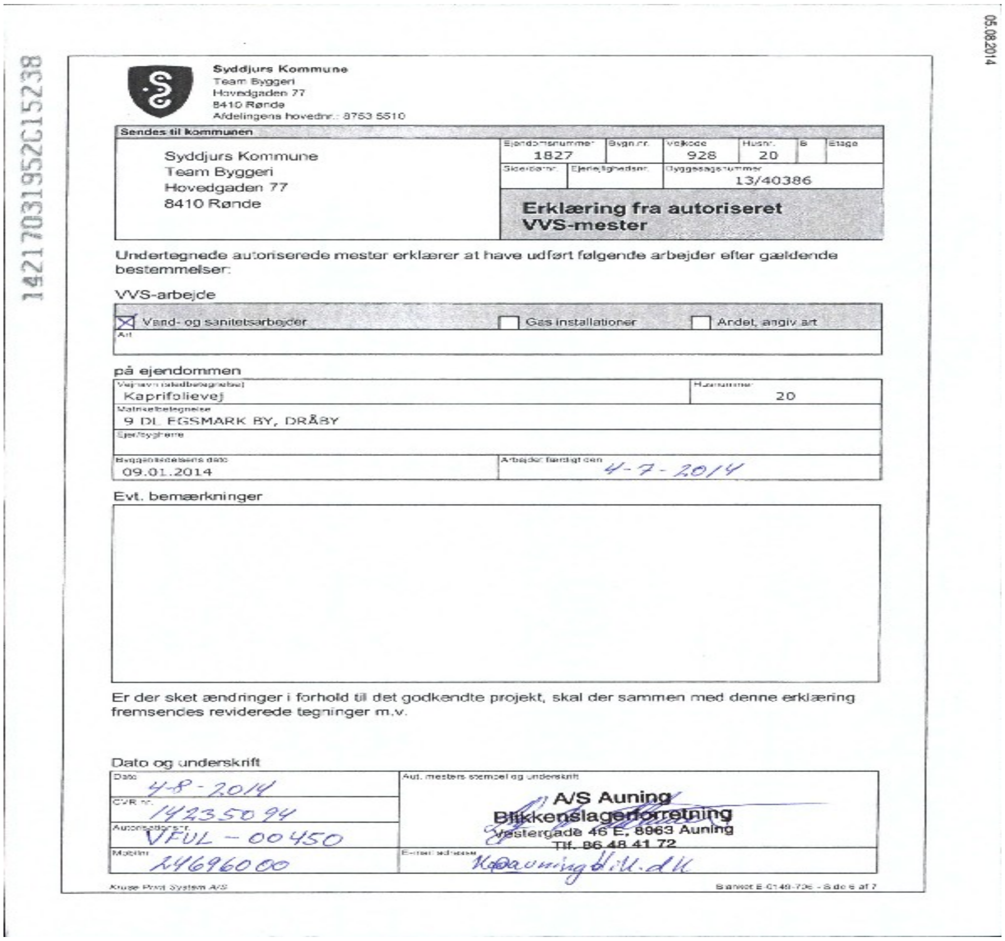
The toolbox

(all is open domain)

Tools (assorted)	***Status 2016***
OCR scanner	OK
De-formating	OK
Tokenizer	OK *
Anonymizer	OK ***
Symbols (num, abbrev, ...)	OK **

(*) = needs localization

Example: Raw document to structured text



14217031952C15238

05.08.2014

Syddjurs Kommune
Team Byggeri
Hovedgaden 77
8410 Rønde
Afdelingens hovednr.: 8753 5510

Sendes til kommunen

Syddjurs Kommune
Team Byggeri
Hovedgaden 77
8410 Rønde

Ejendomsnummer	Bygn.nr.	Vejkode	Husnr.	B	Etage
1827		928	20		

Sidebnr.	Ejerlejlighedsnr.	Byggesagsnummer
		13/40386

Erklæring fra autoriseret VVS-mester

Undertegnede autoriserede mester erklærer at have udført følgende arbejder efter gældende bestemmelser:

VVS-arbejde

☒ Vand- og sanitetsarbejder ☐ Gas installationer ☐ Andet, angiv art

Art

på ejendommen

Vejnavn (stedbetegnelse) Kaprifolievej	Husnummer 20
Matrikelbetegnelse 9 DL EGSMARK BY, DRÅBY	
Ejer/bygherre	
Byggetilladelsens dato 09.01.2014	Arbejdet færdigt den 4-7-2014

Evt. bemærkninger

Er der sket ændringer i forhold til det godkendte projekt, skal der sammen med denne erklæring fremsendes reviderede tegninger m.v.

Dato og underskrift

Dato 4-8-2014	Aut. mesters stempel og underskrift A/S Auning Blikkenslagerforretning Vestergade 46 E, 8963 Auning Tlf. 86 48 41 72
CVR nr. 14235094	
Autorisationsnr. VFUL - 00450	
Mobilnr. 24696000	E-mail adresse Kb@auningtill.dk

Kruse Print System A/S

Blanket E-0149-706 - Side 6 af 7

After OCR scanning

Syddjurs Kommune
Team Byggeri
Hovedgaden 77
8410 Rande
Afdelingens hovednr: 8753 5510

Undertegnede autoriserede mester erklaerer at have udfart falgende arbejder efter gaeldende bestemmelser:

WS-arbejde
VA Vand- ogksanitetsarbejder I Gasinstallationér pé ejendommen

Ejer/bygherre
Byggelilladelsens dato Arbejdet fardigt den 09.01.2014

Preparing anonymization

Syddjurs Kommune
Team Byggeri
Hovedgaden 77
8410 Rande
Afdelingens hovednr: 8753 5510

Undertegnede autoriserede mester erklaerer at have udfart falgende arbejder efter gaeldende bestemmelser:

WS-arbejde
VA Vand- ogksanitetsarbejder I Gasinstallationér pé ejendommen

Ejer/bygherre
Byggelilladelsens dato Arbejdet fardigt den 09.01.2014

Text preparation: Tokenizing, normalizing, anonymizing, ...

Prognose og behandlingsmuligheder**:**
Der er klar **overenstemmelse** mellem de objektive fund og borgers fortælling**..**
- Foreligger der lægeskøn eller udtalelse fra egen læge ang. **pronose**,
arb evne mm?
Ja. Egen læge ser ikke noget arbejdsmarkedsperspektiv for **Marianne**.

---

prognose og behandlingsmuligheder _

der er klar overensstemmelse mellem de objektive fund og borgers fortælling

foreligger der lægeskøn eller udtalelse fra egen læge angående prognose arbejdsevne med mere

ja

egen læge ser ikke noget arbejdsmarkedsperspektiv for _FirstName_

Second action point: Preparing generic specs and documents

Documents available in generic/embryonic forms:

Databehandleraftaler ("data processing agreement")
Systemkrav ("system requirements", specs)
Udbudsmaterialer ("bidding materials")
...

Third action point: Managing the bidding situation

Bidding material should specify

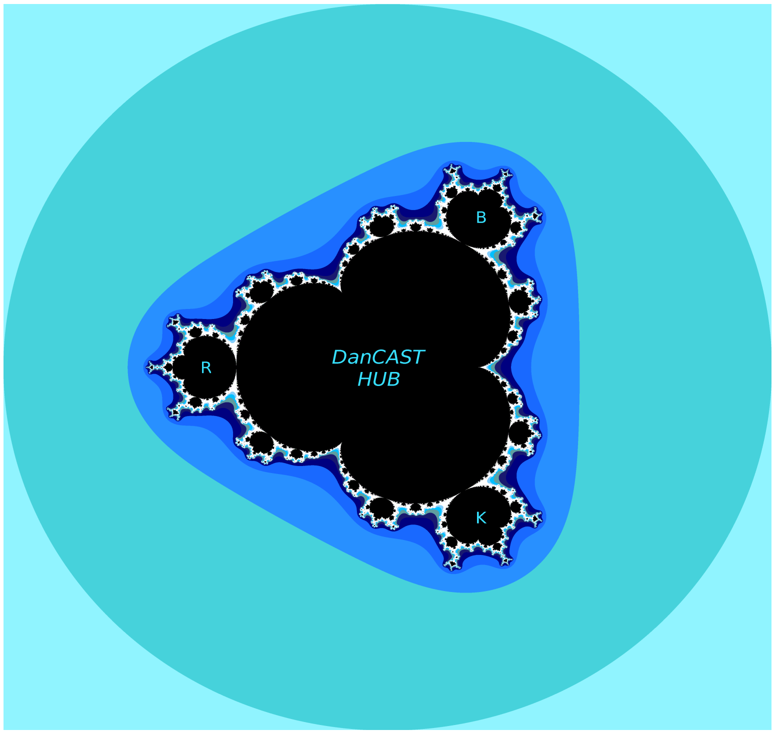
- clear divison of components (dictionary, AM, LM)
- transparent recycling (reused vs. new data)
- only latest-state of databases can be owned by provider
- short update-cycle for AM and LM

Either as requirements or as desiderata

Bidding material should refer to the shared corpora, e.g. qualification round

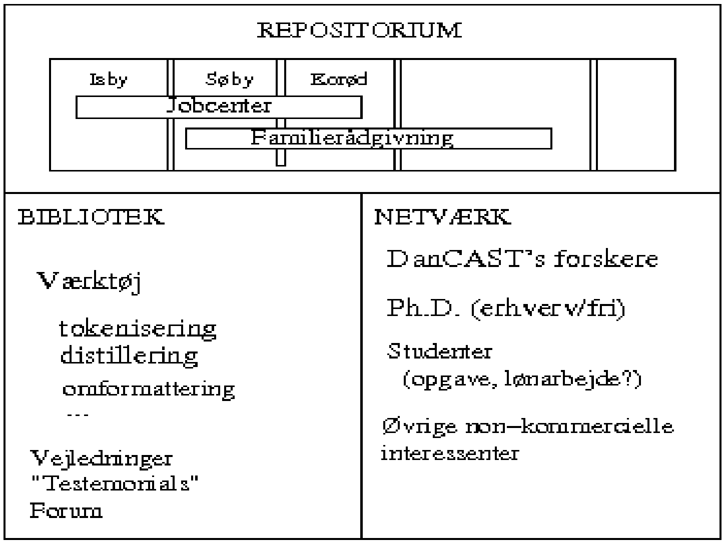
These features are as important for **return-of-investment** as are WER etc.

The HUB



www.dancast.dk

The DanCAST hub for ASR localization support



REPOSITORIUM

Isby

Søby

Korød

Jobcenter

Familierådgivning

BIBLIOTEK

Værktøj

tokenisering
distillering
omformattering
...

Vejledninger
"Testemonials"
Forum

NETVÆRK

DanCAST's forskere

Ph.D. (erhverv/fri)

Studenter
(opgave, lønarbejde?)

Øvrige non–kommercielle interessenter

Status quo 2016

May 2015:	First 100% Danish-Danish contract
Januar 2016:	First independent SME established
July 2016:	Now 3 Danish SME start-ups

Neither would have existed without the shared HUB-data

Bidding rounds are now more fair (results not given in advance)

Economically most significant result:

> Licence conditions are vastly improved,
> even in *old* contracts!

So, a happy end?

Not entirely:

Government cuts in 2015 and 2016 in all public budgets
Also universities are currently firing researchers

However:

The HUB survives and is now almost self-sustaining

By way of conclusion

If we could start over...

Step 1. Prepare the ground

IT responsibles: Nuts-and-bolts courses
Decision makers: Adjusted expectations!
End-users: interest groups for employees

Step 2. Collect existing materials

Corpora
Tools for annotation, alignment, tagging, anonymization, ...

Step 3. Establish a data portal

Restricted entry - for individual municipalities
Semi-restricted entry - data-sharing among municipalities
Unrestricted entry - fully processed and anonymized data

Step 4. Create a library of generic formulas (semi-restricted access)

Agreements, tender material, specs

Step 5. Prepare for bidding rounds

Organize groups of municipalities
Require!

THE END