Enhanced Immersivity: Using Speech Recognition for More Natural Player AI Interactions

Gautier Boeda
AI Engineer – SQUARE ENIX CO., LTD

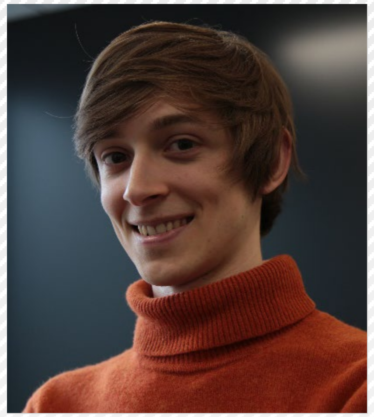
BEFORE GOING FURTHER

- This is an experimental project
  - Still in R&D: still unproven in a real game

- The contents shown today has been created for conferences and studies purposes
  - It is not a new IP.

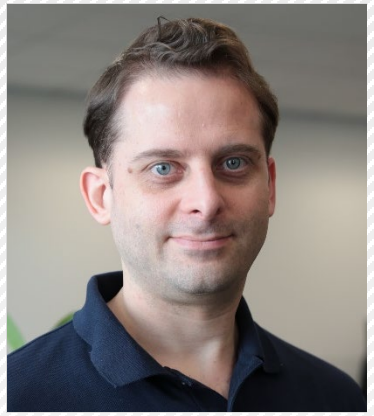
TEAM

SQUARE ENIX JAPAN – ADVANCED TECHNOLOGY DIVISION



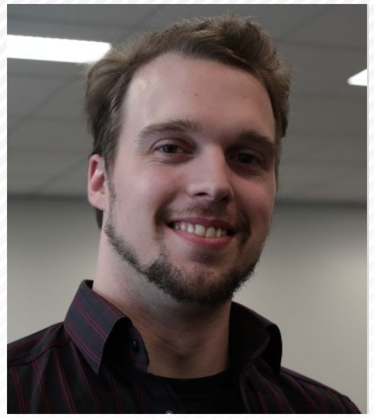
Gautier Boeda

Yuta Mizuno



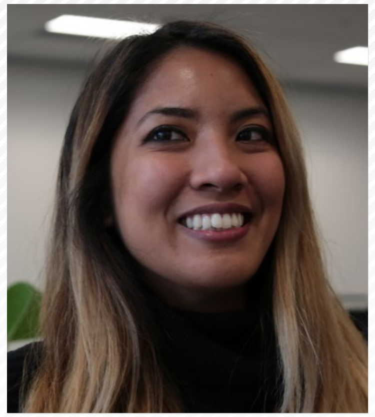
Remi Driancourt

Brian Wanamaker



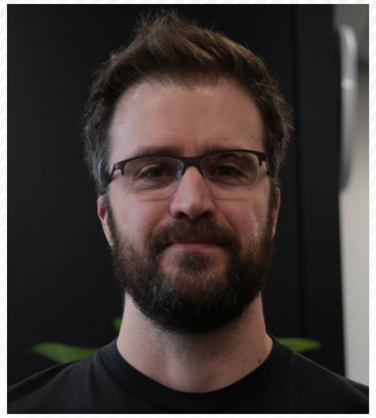
Perry Leijten

Stephanie Timmins



Adelle Bueno

Eduardo Mosena



Louis-Philippe Sanschagrin

MOTIVATION

WHAT ARE WE TRYING TO IMPROVE?

- Non-playable characters in virtual reality feel really close!
  - Enhance immersion
- Interacting with them felt sloppy, breaking the immersion
  - Limited to buttons or other classic mechanism, giving a sensation of being a ghost.

MOTIVATION

HOW CAN IT BE ACHIEVED?

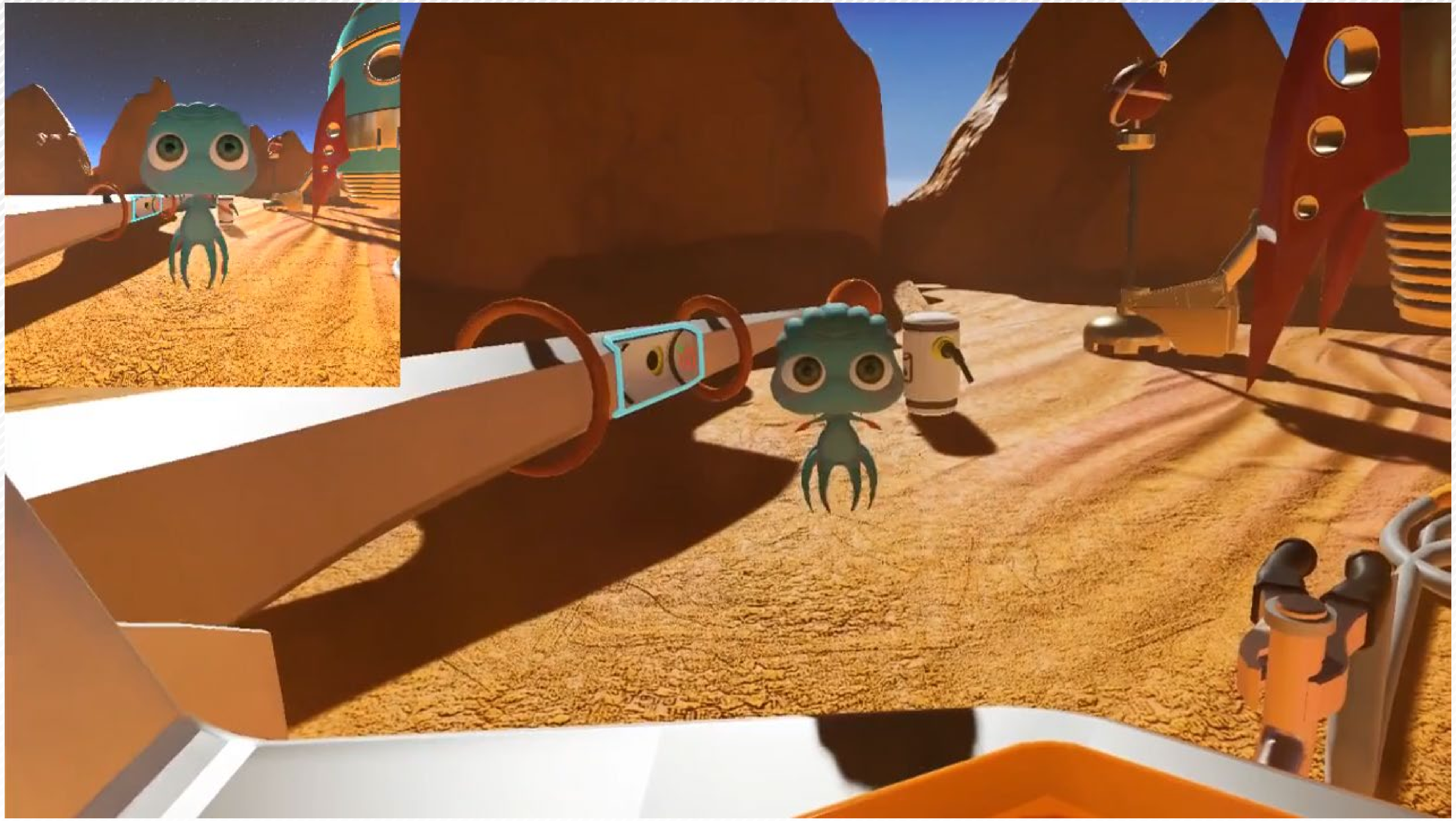
- Mission
  - Bring more natural interactions:
    - Voice interaction
    - Body interaction

So that the agent can understand

- Where we currently are
- What we are talking about
- Where we are pointing at
- Where we are looking at
- What we are currently doing

DEMO

FIRST GLANCE AT KOBUN



[View Video (Click)]

WHAT'S ON THE MENU TODAY?

- Speech recognition pipeline
  - Pipeline explanation
  - Failure cases
    - With their solutions

- Interactions
  - Pointing at location while giving instructions
  - Location-based information disambiguation

SPEECH RECOGNITION PIPELINE

PIPELINE SUMMARY

Speech Recognition

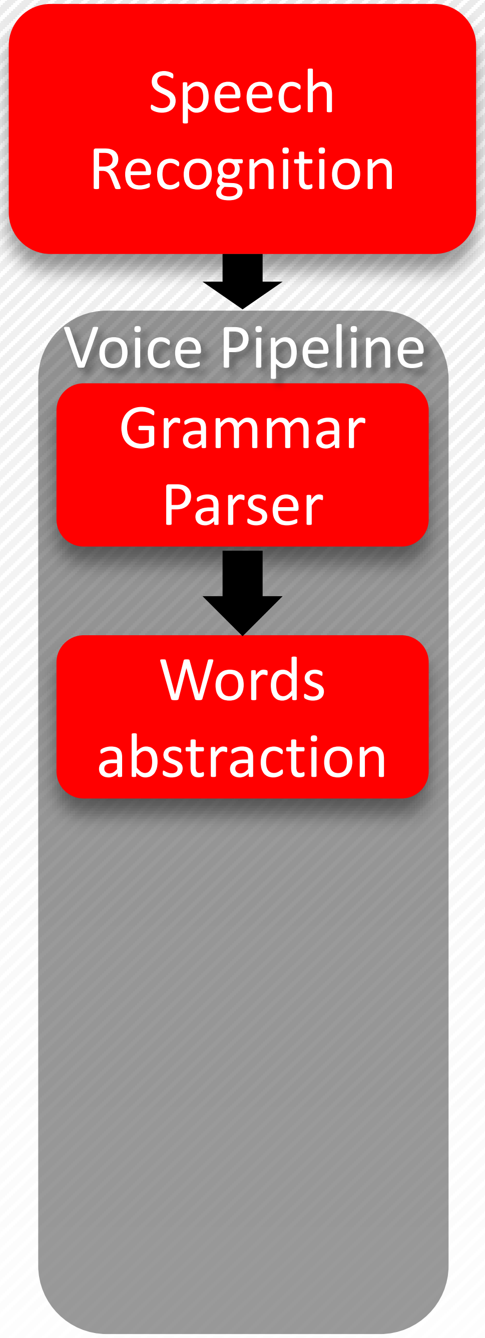
Pick up an enormous apple
[Verb: Pick] [Preposition: up] [Determiner: an] [Adjective: enormous] [Noun: apple]

Engine: Julius (github.com/julius-speech/julius)

- Real-time
- Word timestamp
  - Useful for linking the voice to the actions of the speaker:
    - “Go there!” -> “there” was said 0.84s ago.
    - Where was pointing the player 0.84s ago? -> Vector3(x, y, z)
- Support any language (need to provide the model)
  - Japanese model: very good
    - Diverse audience, some accents
    - Provide part-of-speech

SPEECH RECOGNITION PIPELINE

PIPELINE SUMMARY



Pick up an enormous apple

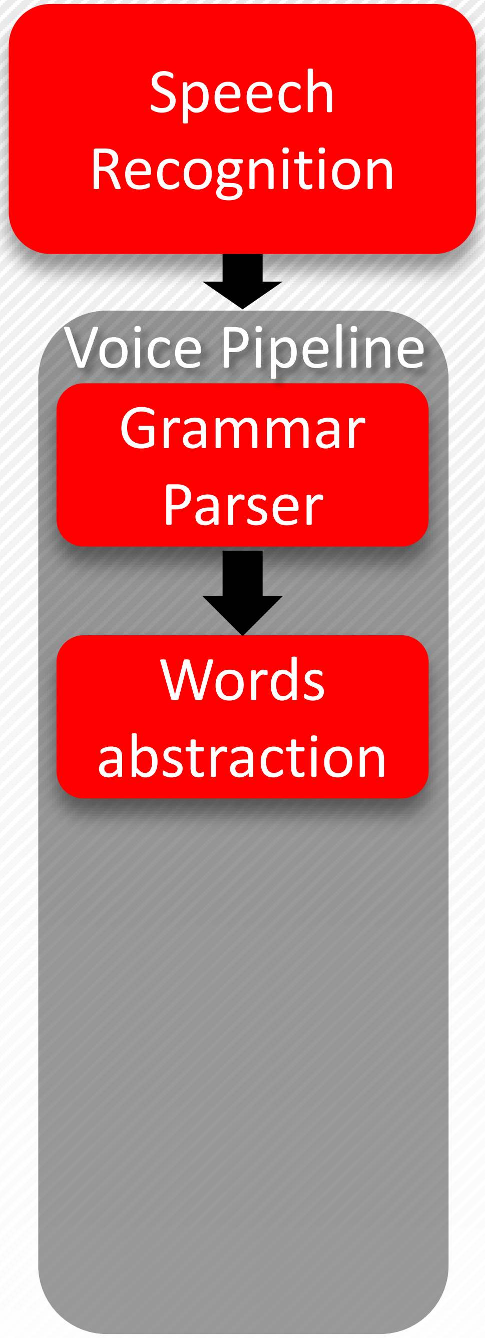
[Verb: Pick] [Preposition: up] [Determiner: an] [Adjective: enormous] [Noun: apple]

[Verb: Pick up] [Predicate: enormous] [Object: apple]

[Verb: 🟩🟨🟩🟥] [Predicate: 🟥🟪🟦] [Object: 🟪🟪⬜🟥🟪]

SPEECH RECOGNITION PIPELINE

WORDS ABSTRACTION

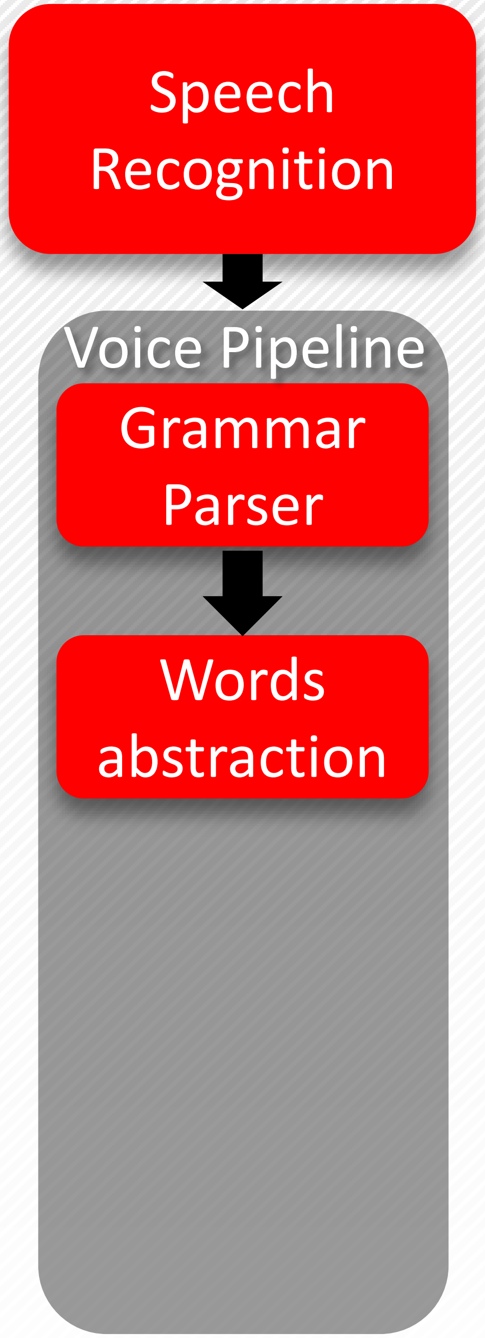


- Problem to solve:
  - Support multiple languages without limiting the player’s set of vocabulary
- Cause of the Problem:
  - Words are language-based. They don’t have bindings between languages.

    We need to abstract them.
- Idea:
  - Can we create the DNA of a word? What could be the genes?

SPEECH RECOGNITION PIPELINE

WORDS ABSTRACTION

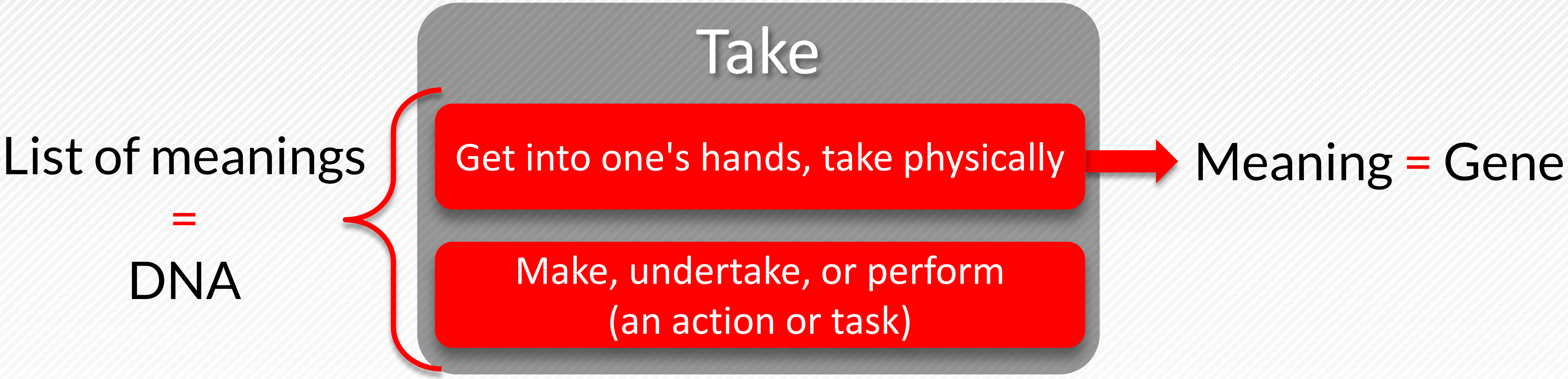


Take an apple

Get into one's hands, take physically

Take a break

Make, undertake, or perform (an action or task).



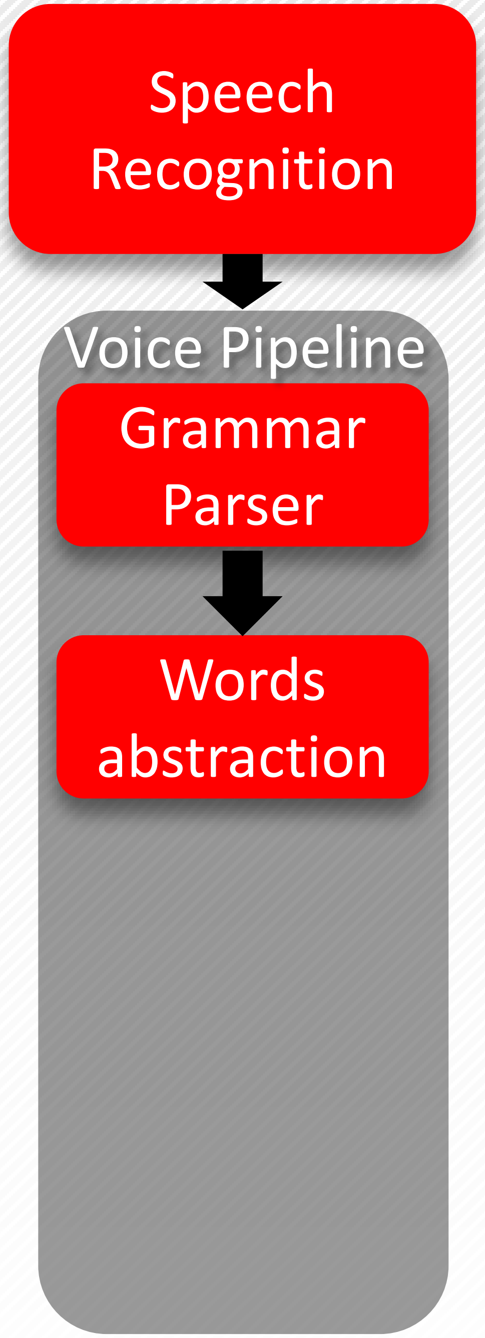
List of meanings
=
DNA
Take
Get into one's hands, take physically
Make, undertake, or perform (an action or task)
Meaning = Gene

How? → WordNet

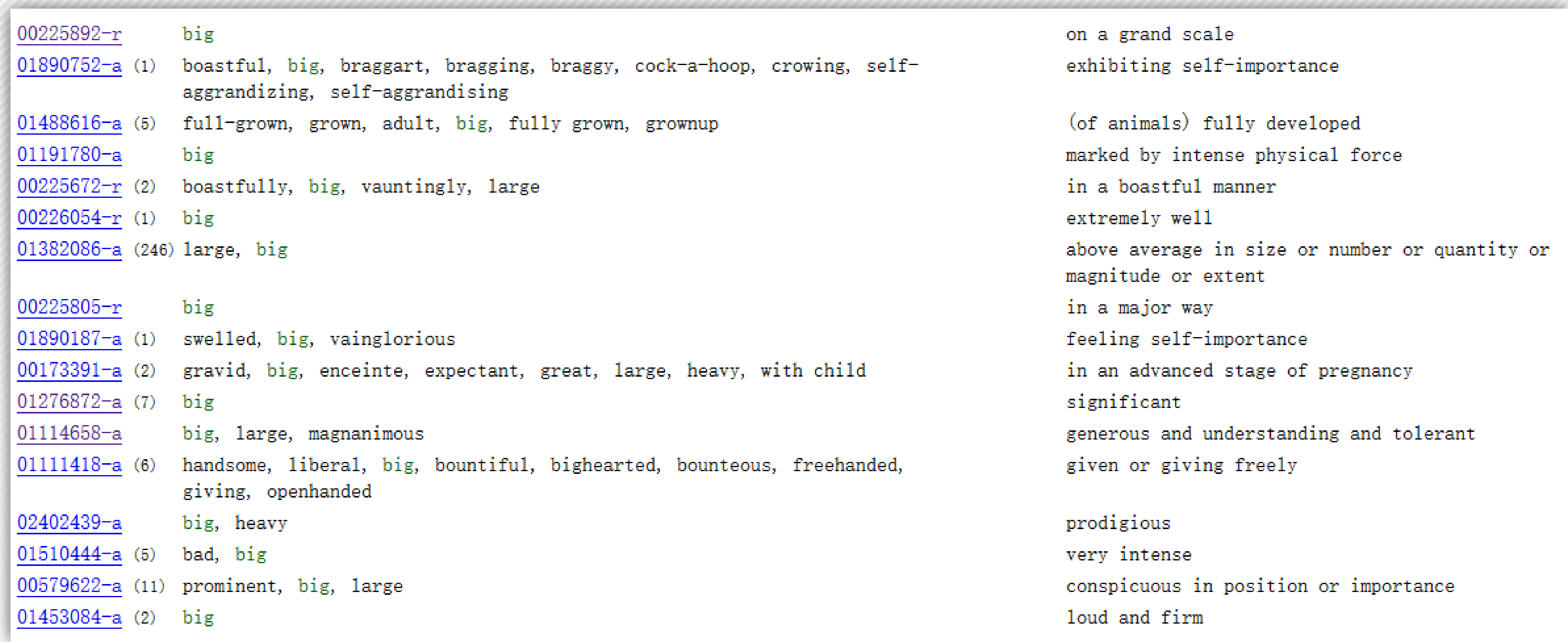
- Database of "sets of cognitive synonyms (synset), each expressing a distinct concept" wordnet.princeton.edu/
- Support multiple languages

SPEECH RECOGNITION PIPELINE

WORDS ABSTRACTION



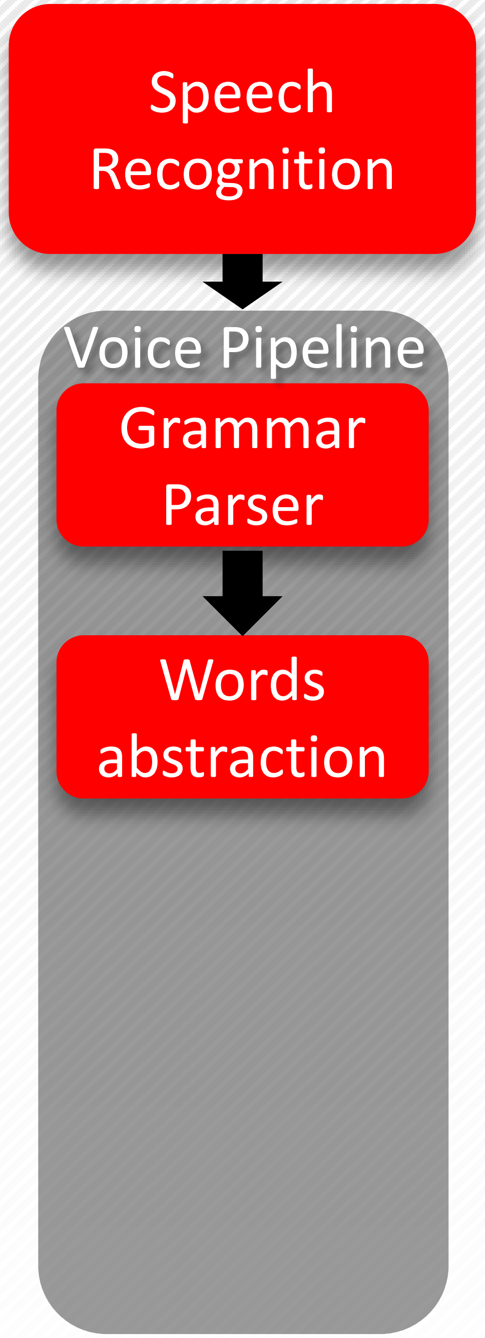
- Example:
  - We need a concept of “Big” in our experience, as in “A big apple”



00225892-r		big	on a grand scale
01890752-a	(1)	boastful, big, braggart, bragging, braggy, cock-a-hoop, crowing, self-aggrandizing, self-aggrandising	exhibiting self-importance
01488616-a	(5)	full-grown, grown, adult, big, fully grown, grownup	(of animals) fully developed
01191780-a		big	marked by intense physical force
00225672-r	(2)	boastfully, big, vauntingly, large	in a boastful manner
00226054-r	(1)	big	extremely well
01382086-a	(246)	large, big	above average in size or number or quantity or magnitude or extent
00225805-r		big	in a major way
01890187-a	(1)	swelled, big, vainglorious	feeling self-importance
00173391-a	(2)	gravid, big, enceinte, expectant, great, large, heavy, with child	in an advanced stage of pregnancy
01276872-a	(7)	big	significant
01114658-a		big, large, magnanimous	generous and understanding and tolerant
01111418-a	(6)	handsome, liberal, big, bountiful, bighearted, bounteous, freehanded, giving, openhanded	given or giving freely
02402439-a		big, heavy	prodigious
01510444-a	(5)	bad, big	very intense
00579622-a	(11)	prominent, big, large	conspicuous in position or importance
01453084-a	(2)	big	loud and firm

SPEECH RECOGNITION PIPELINE

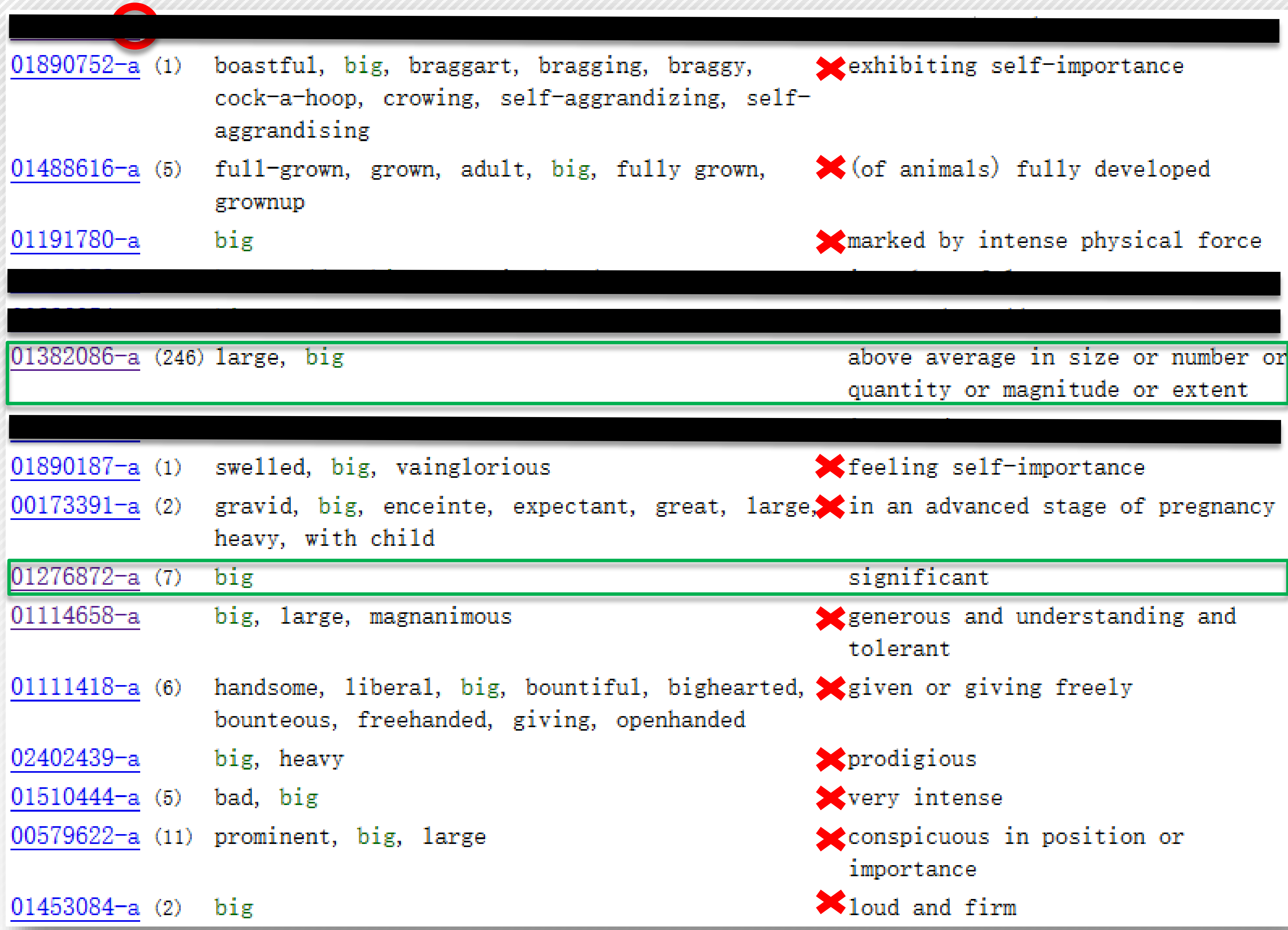
WORDS ABSTRACTION



Which “big” meaning are we interested in?

1) Keep adjectives
   r = adverb
   a = adjective

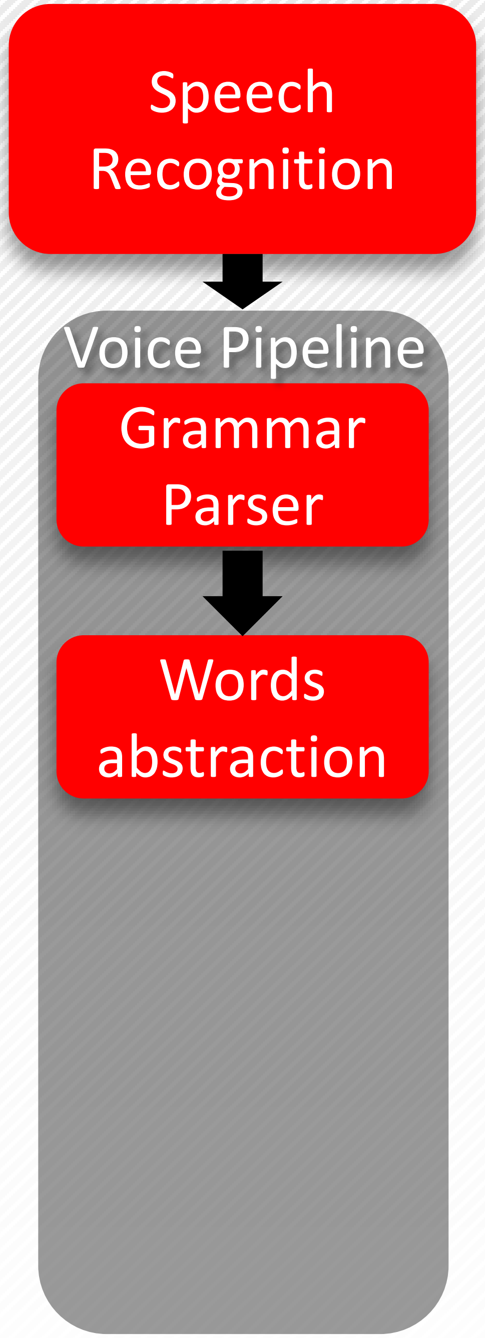
2) Select concepts



[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
01890752-a	(1)	boastful, big, braggart, bragging, braggy, cock-a-hoop, crowing, self-aggrandizing, self-aggrandising	✗	exhibiting self-importance
01488616-a	(5)	full-grown, grown, adult, big, fully grown, grownup	✗	(of animals) fully developed
01191780-a		big	✗	marked by intense physical force
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
01382086-a	(246)	large, big		above average in size or number or quantity or magnitude or extent
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
01890187-a	(1)	swelled, big, vainglorious	✗	feeling self-importance
00173391-a	(2)	gravid, big, enceinte, expectant, great, large, heavy, with child	✗	in an advanced stage of pregnancy
01276872-a	(7)	big		significant
01114658-a		big, large, magnanimous	✗	generous and understanding and tolerant
01111418-a	(6)	handsome, liberal, big, bountiful, bighearted, bounteous, freehanded, giving, openhanded	✗	given or giving freely
02402439-a		big, heavy	✗	prodigious
01510444-a	(5)	bad, big	✗	very intense
00579622-a	(11)	prominent, big, large	✗	conspicuous in position or importance
01453084-a	(2)	big	✗	loud and firm

SPEECH RECOGNITION PIPELINE

WORDS ABSTRACTION



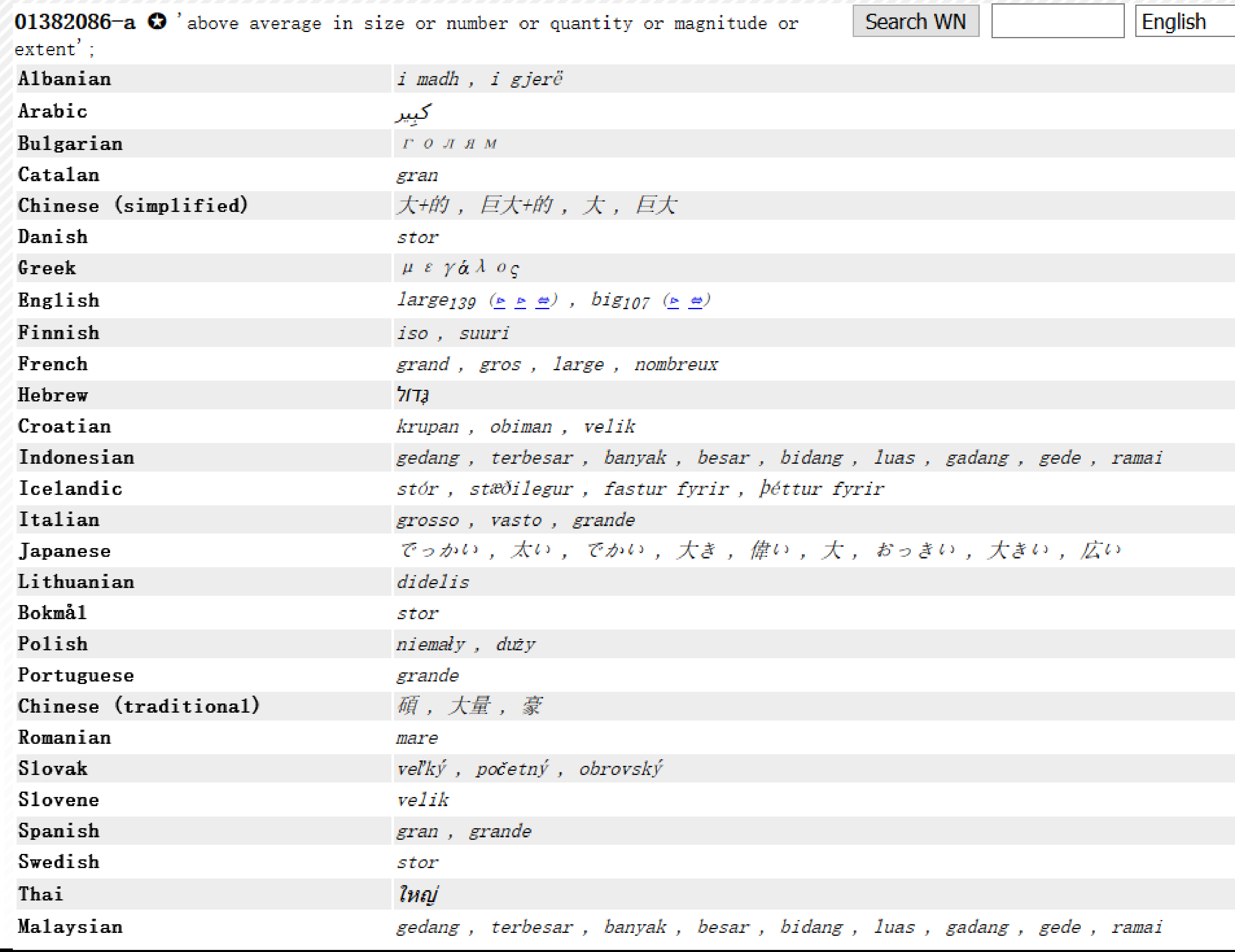
- Our “Big” predicate DNA will be composed of:

  [01382086-a] above average in size or number or quantity or magnitude or extent

  [01276872-a] Significant

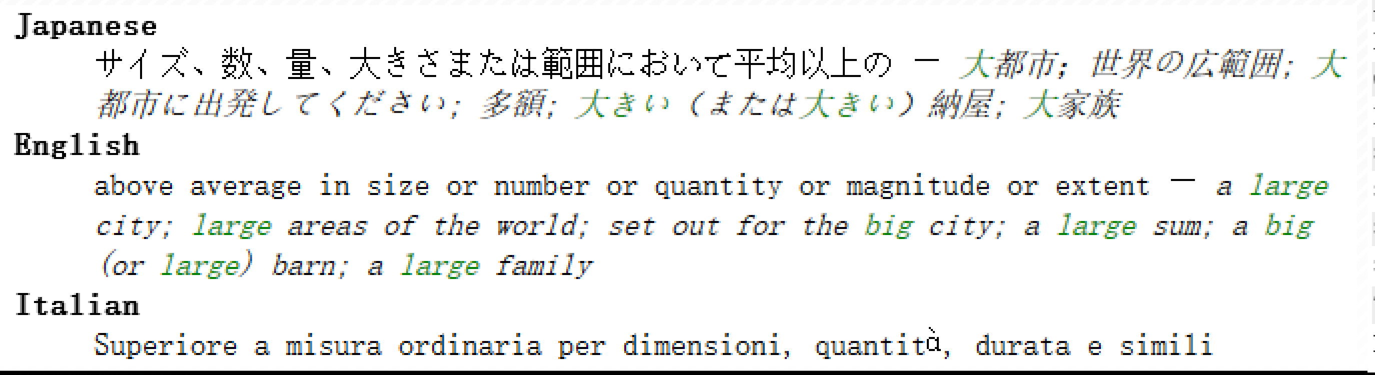
Big 🟥

- Check our synsets:
  - Multi languages!



01382086-a ✪ 'above average in size or number or quantity or magnitude or extent';

Albanian	i madh , i gjerë
Arabic	كبير
Bulgarian	голям
Catalan	gran
Chinese (simplified)	大+的 , 巨大+的 , 大 , 巨大
Danish	stor
Greek	μεγάλος
English	large$_{139}$ (▻ ▻ ⇔) , big$_{107}$ (▻ ⇔)
Finnish	iso , suuri
French	grand , gros , large , nombreux
Hebrew	גָּדוֹל
Croatian	krupan , obiman , velik
Indonesian	gedang , terbesar , banyak , besar , bidang , luas , gadang , gede , ramai
Icelandic	stór , stæðilegur , fastur fyrir , þéttur fyrir
Italian	grosso , vasto , grande
Japanese	でっかい , 太い , でかい , 大き , 偉い , 大 , おっきい , 大きい , 広い
Lithuanian	didelis
Bokmål	stor
Polish	niemały , duży
Portuguese	grande
Chinese (traditional)	碩 , 大量 , 豪
Romanian	mare
Slovak	veľký , početný , obrovský
Slovene	velik
Spanish	gran , grande
Swedish	stor
Thai	ใหญ่
Malaysian	gedang , terbesar , banyak , besar , bidang , luas , gadang , gede , ramai



Japanese

サイズ、数、量、大きさまたは範囲において平均以上の ー *大都市; 世界の広範囲; 大都市に出発してください; 多額; 大きい（または大きい）納屋; 大家族*

English

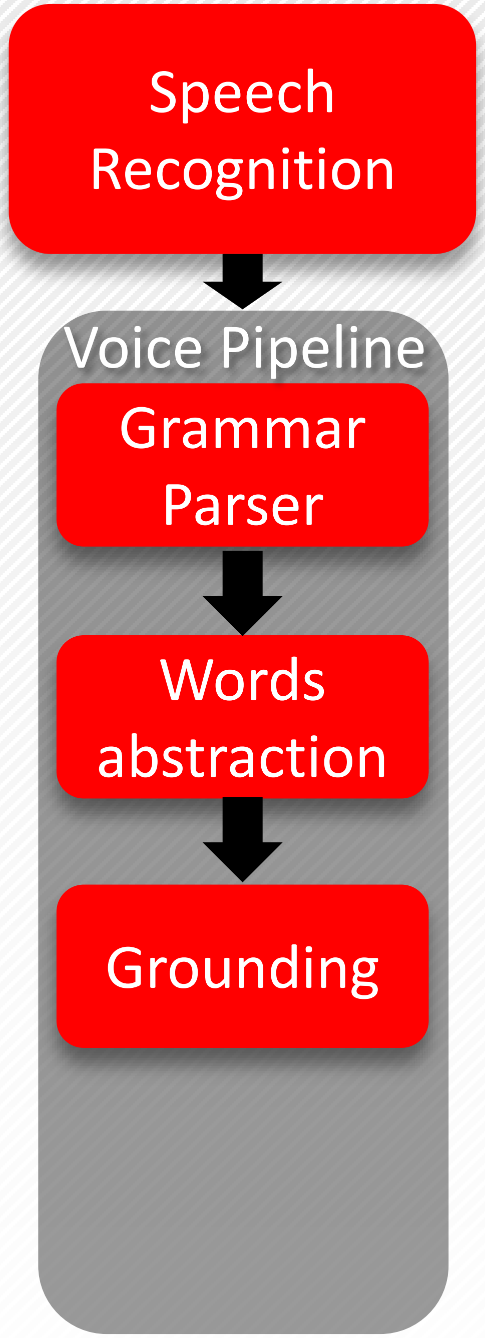
above average in size or number or quantity or magnitude or extent ー *a large city; large areas of the world; set out for the big city; a large sum; a big (or large) barn; a large family*

Italian

Superiore a misura ordinaria per dimensioni, quantità, durata e simili

SPEECH RECOGNITION PIPELINE

PIPELINE SUMMARY



Pick up an enormous apple
[Verb: Pick] [Preposition: up] [Determiner: an] [Adjective: enormous] [Noun: apple]

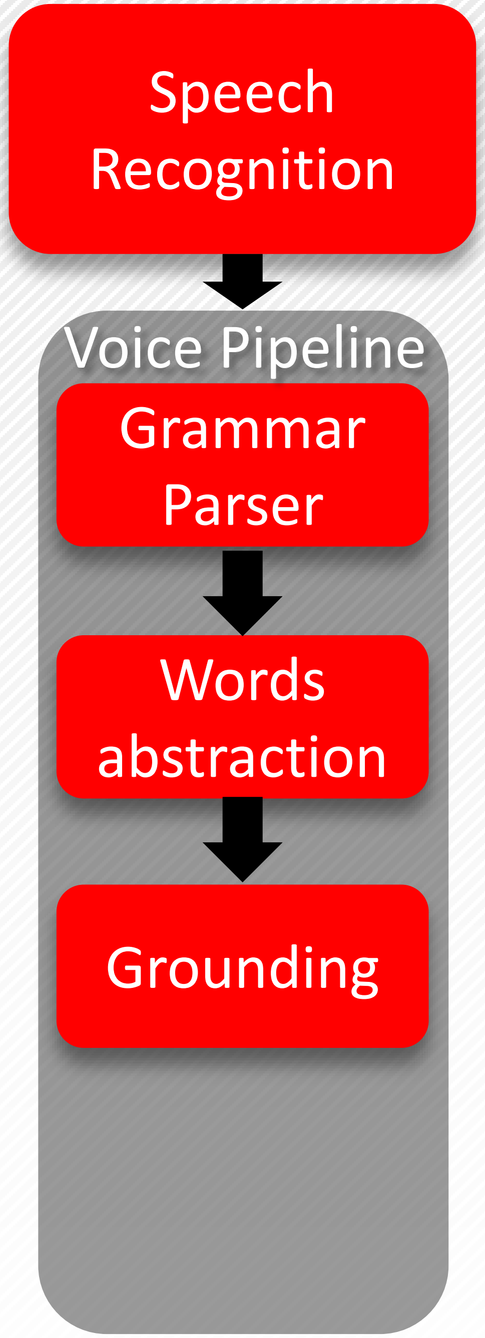
[Verb: Pick up] [Predicate: enormous] [Object: apple]

[Verb: 🟩🟧🟩🟥] [Predicate: 🟥🟪🟦] [Object: 🟪🟪⬜🟥🟪]

[Take] [big] [apple]

SPEECH RECOGNITION PIPELINE

GROUND THE WORDS INTO THE CONCEPTS OF OUR WORLD



Speech Recognition

Voice Pipeline

Grammar Parser

Words abstraction

Grounding

Ground the abstracted words to our concepts:

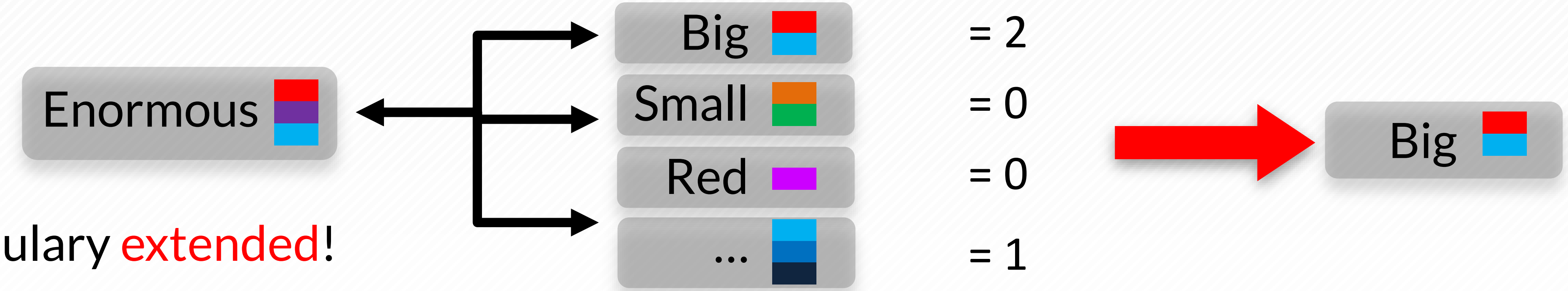
- Locations (above, behind, left, etc)
- Predicates (color, size, etc)
- Verbs
- …

Using a utility-based scoring method.

Example:

Word to ground (DNA)

Predicates (Concepts of our world)



Enormous 🟥🟪🟦

Big 🟥🟦 = 2

Small 🟧🟩 = 0

Red 🟪 = 0

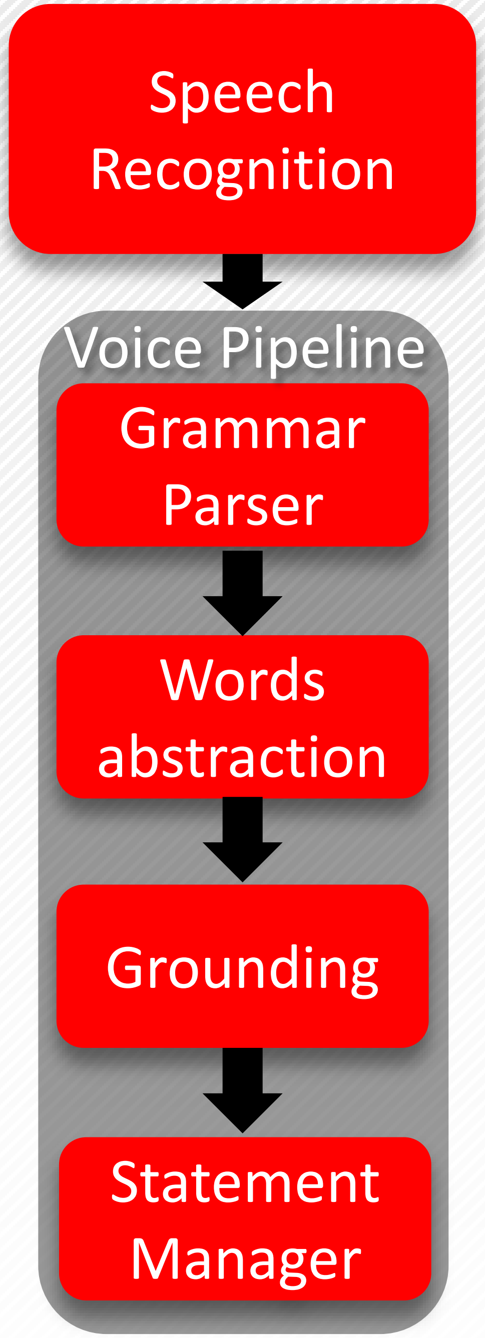
… 🟦🟦⬛ = 1

Big 🟥🟦

Player's set of vocabulary extended!

SPEECH RECOGNITION PIPELINE

PIPELINE SUMMARY



Speech Recognition

Voice Pipeline

Grammar Parser

Words abstraction

Grounding

Statement Manager

Pick up an enormous apple

[Verb: Pick] [Preposition: up] [Determiner: an] [Adjective: enormous] [Noun: apple]

[Verb: Pick up] [Predicate: enormous] [Object: apple]

[Verb: 🟩🟧🟩🟥] [Predicate: 🟥🟪🟦] [Object: 🟪🟪⬜🟥🟪]

[Take] [big] [apple]

Store the statement in memory. [Take] [big] [apple]

FAILURE CASES

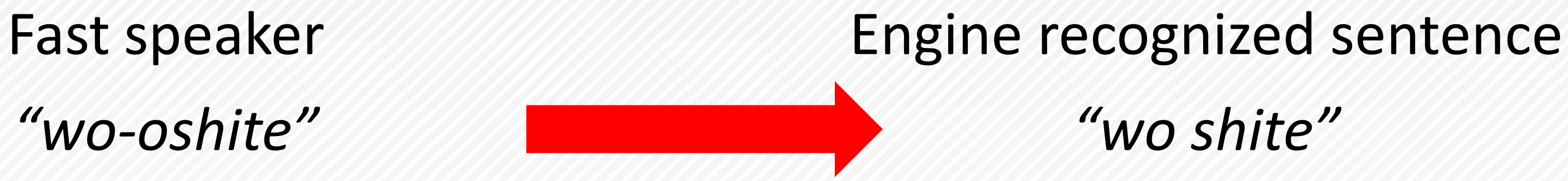
- Connection of words
- Homonyms
- Longer sentences take longer to parse, disturbing the player

FAILURE CASES

CONNECTION OF WORDS

- “wo shite” and “wo oshite”

- Fast speaker will link “wo” and “oshite”.



Fast speaker

“wo-oshite”

Engine recognized sentence

“wo shite”

Solution:

- Addition of a layer of context-based translation.
  - However, it is not perfect.

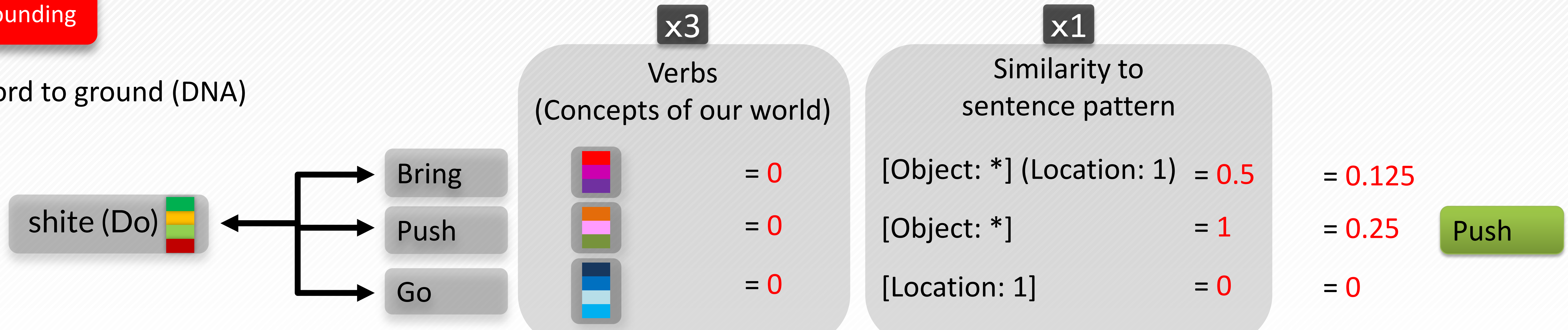
FAILURE CASES

CONNECTION OF WORDS

Fast speaker		Engine recognized sentence
“wo-oshite” (Push)	→	*“wo shite” (Do)*

Speech Recognition	[Noun: botan (button)] [Particle: wo] [Verb: shite (Do)]
Grammar Parser	[Verb: shite (Do)] [Object: button]
Words abstraction	[Verb: 🟩🟨🟩🟥] [Object: 🟪🟦]
Grounding	

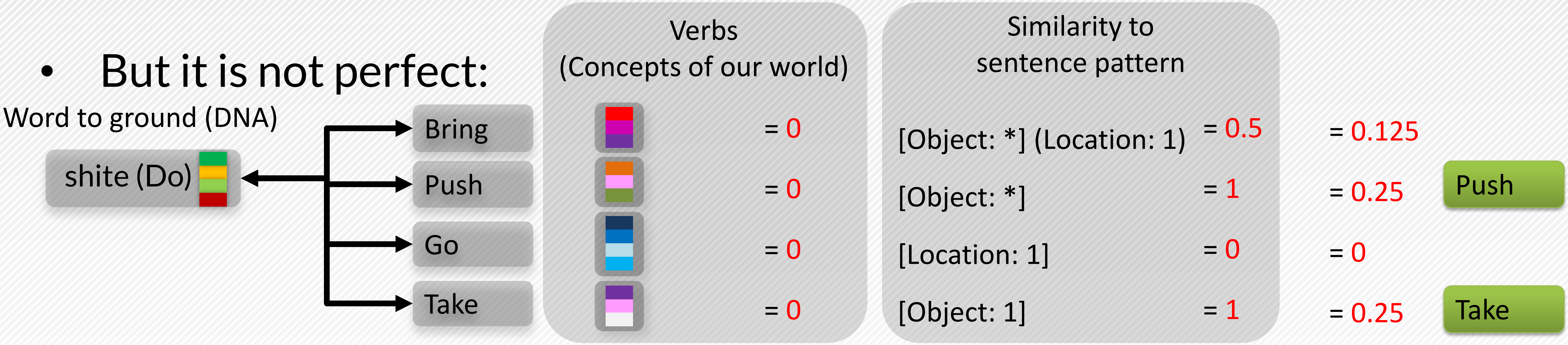
Word to ground (DNA)



FAILURE CASES

CONNECTION OF WORDS

- But it is not perfect:



Word to ground (DNA)

shite (Do)

Verbs
(Concepts of our world)

Bring = 0

Push = 0

Go = 0

Take = 0

Similarity to
sentence pattern

[Object: *] (Location: 1) = 0.5 = 0.125

[Object: *] = 1 = 0.25 Push

[Location: 1] = 0 = 0

[Object: 1] = 1 = 0.25 Take

- If the engine provides it: Use one of the other sentence candidate.
- If still not enough: Pronunciation similarity in the given language

FAILURE CASES

HOMONYMS

Verb: “hanasu” can be spelled:

- 話す = to speak
- 離す = to separate
- 放す = to release

By lack of context (not aware of our world), the engine can make a mistake.

FAILURE CASES

HOMONYMS

Solution:

[Verb: 話す(speak)] [Pronoun: it]

1. Translate the verbs into their pronunciation (話す → hanasu)

   [Verb: hanasu] [Object: it]

2. Find all verbs with the same pronunciation

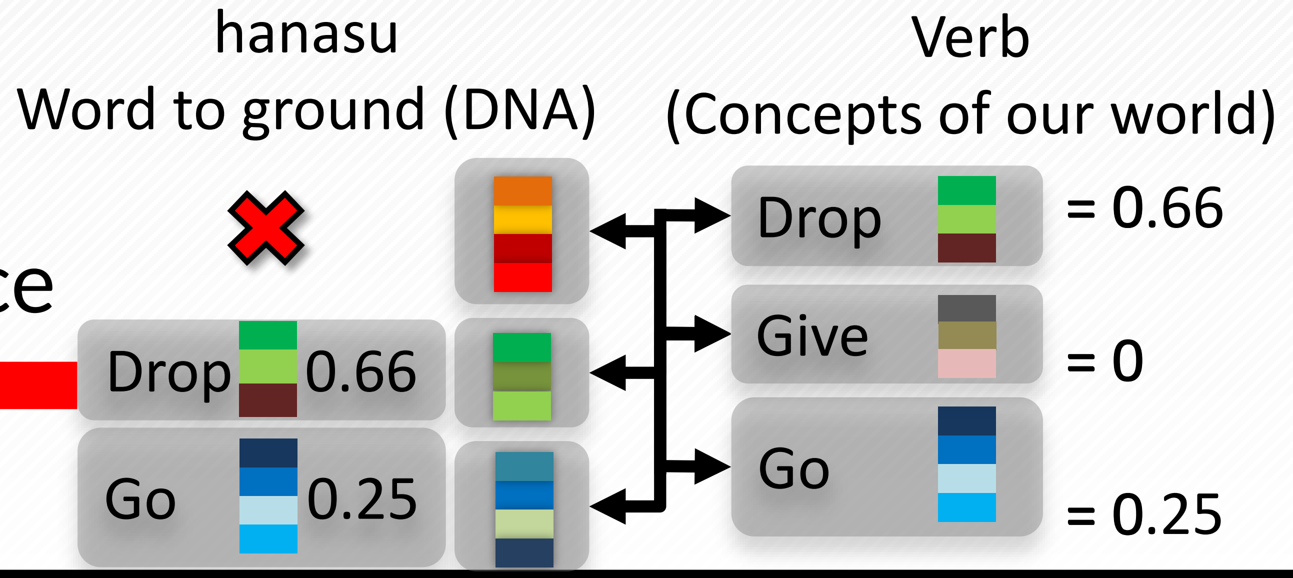
   [Verb: hanasu(話す(speak), 放す(release), 離す(separate))] [Object: it]

3. Abstract these verbs into concepts

   [Verb: hanasu(🟧🟨🟥🟥, 🟩🟩🟩, 🟦🟦🟩🟦)] [Object: it]

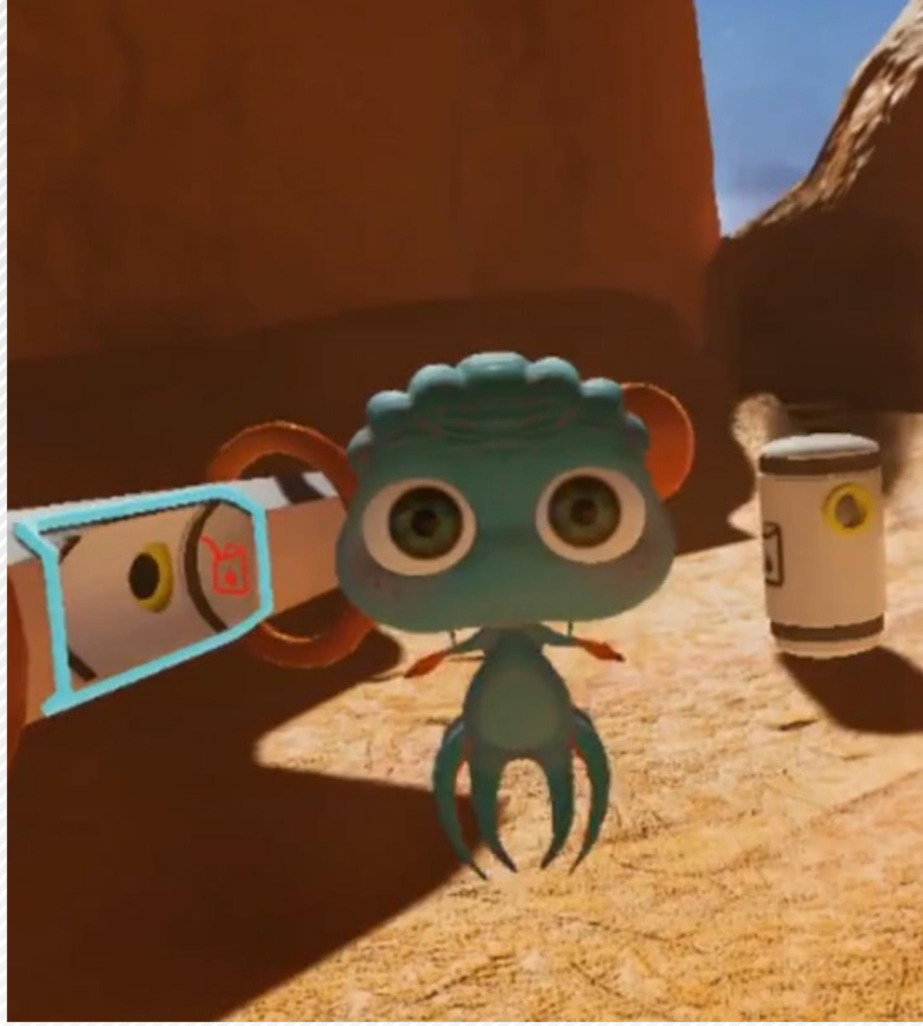
4. Compare them with the concepts of our experience

   [Drop] [Object: it]



FAILURE CASES

LONGER SENTENCES TAKE LONGER TO PARSE



- User becomes uncomfortable.

Solution:

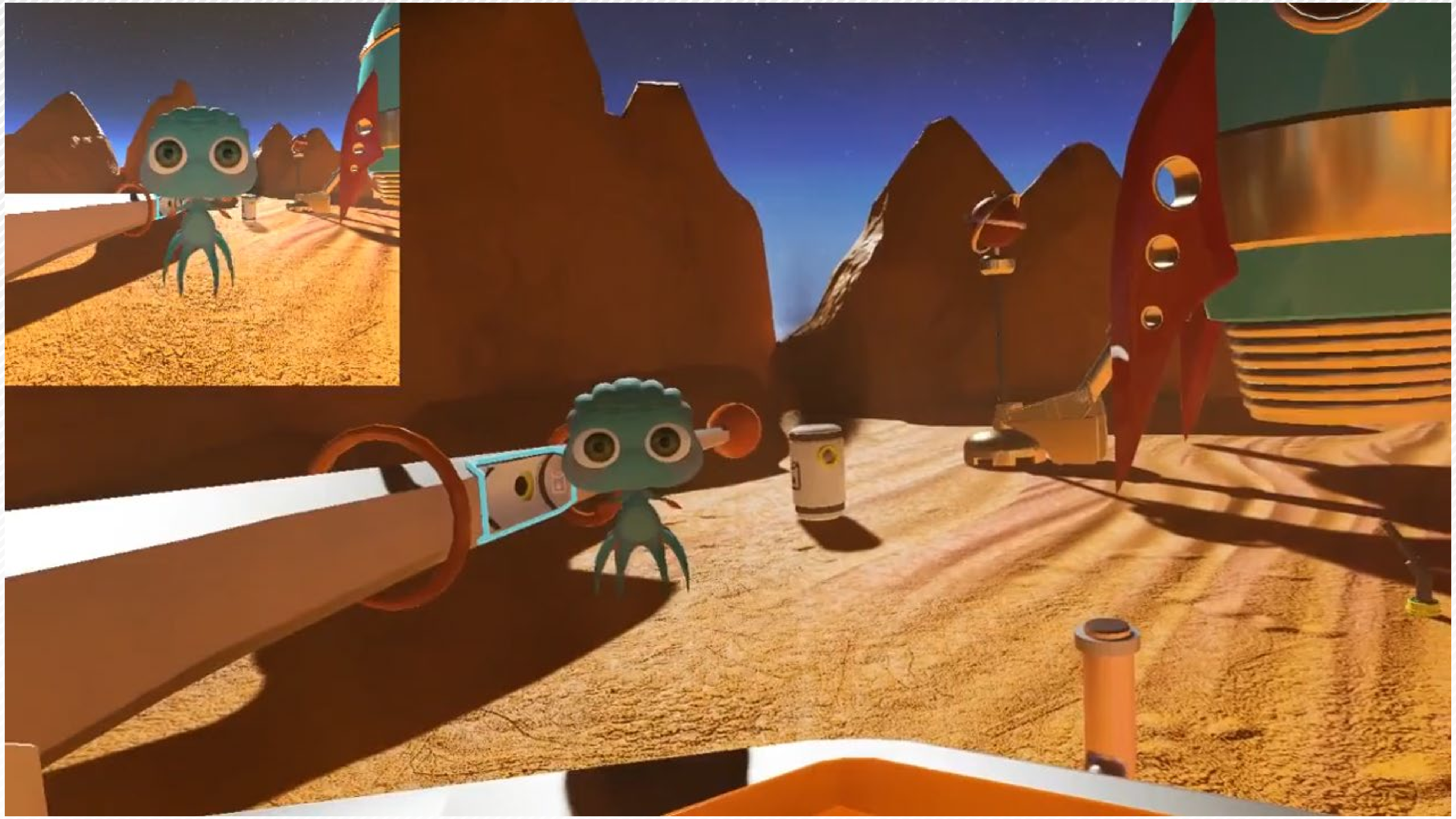
- Add feedback to the AI agent:
  - “Thinking” posture
  - I did not understand your speech
  - I did not find what you were talking about
  - I understood but I don’t have the ability to execute your request
  - I don’t like you, therefore I won’t listen to you
  - I don’t like the object, therefore I won’t execute your request

INTERACTIONS

- Pointing at location, objects while giving instructions
  - Go there
  - Bring me this apple
- Location-based information disambiguation
  - Go on the left of the table
  - Take the apple that is behind the TV

POINTING AT LOCATION, OBJECTS

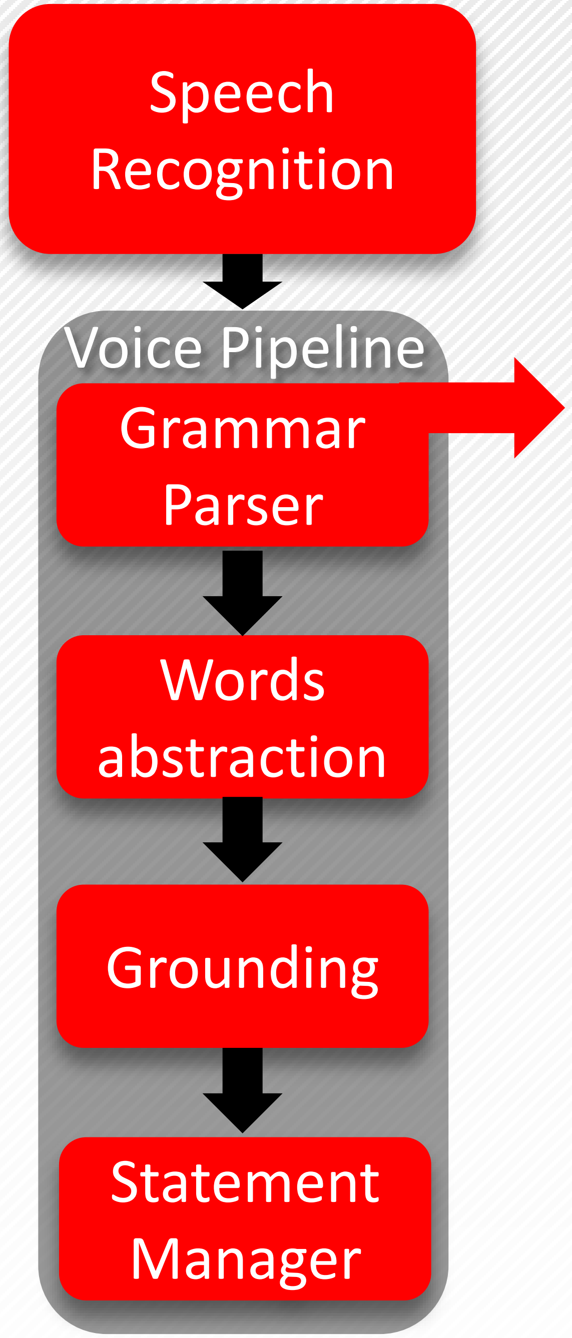
EXAMPLE



[View Video (Click)]

POINTING AT LOCATION, OBJECTS

GRAMMAR



Speech Recognition

Voice Pipeline

Grammar Parser

Words abstraction

Grounding

Statement Manager

- Take this apple.
  - This/that → Absolute Determiner

[Verb: Take] [AbsoluteDeterminer: this] [Object: apple]

- Bring that.
  - This/that → Absolute Object

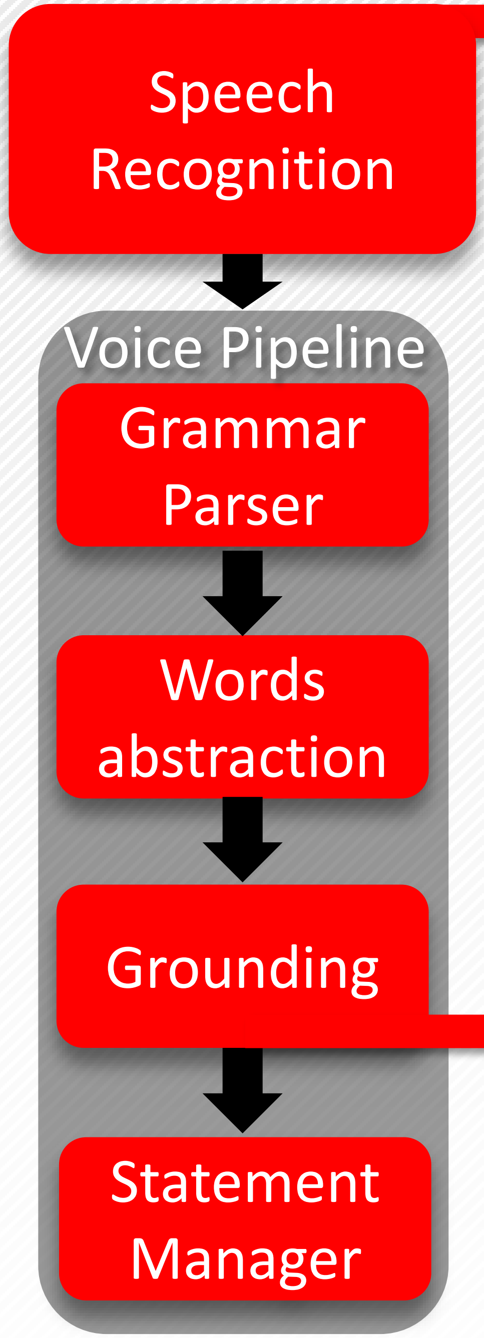
[Verb: Bring] [AbsoluteObject: this]

- Go there.
  - There → Absolute Location

[Verb: Go] [AbsoluteLocation: there]

POINTING AT LOCATION, OBJECTS

GROUNDING



Julius: Speech recognition engine

- Word timestamp

Time between [Word has been said] ←→ [Sentence has been parsed]

Go there

[Verb: Go: 0.89] [Preposition: there: 0.53]

[Verb: Go] [AbsoluteLocation: 0.53]

[Verb: ■] [AbsoluteLocation: 0.53]

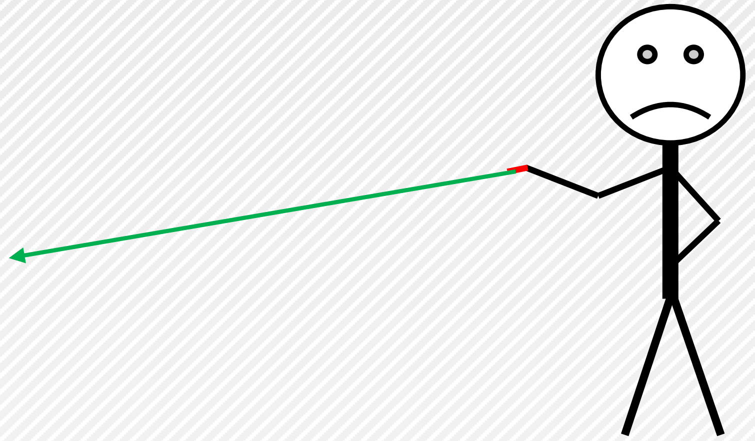
[Go] [AbsoluteLocation: Pointing Information (Position, Direction)]

Ground the absolute location-based information

How? → Get the player’s pointing information (when the word was said)

POINTING AT LOCATION, OBJECTS

POINTING INFORMATION: NATURAL POINTING METHOD

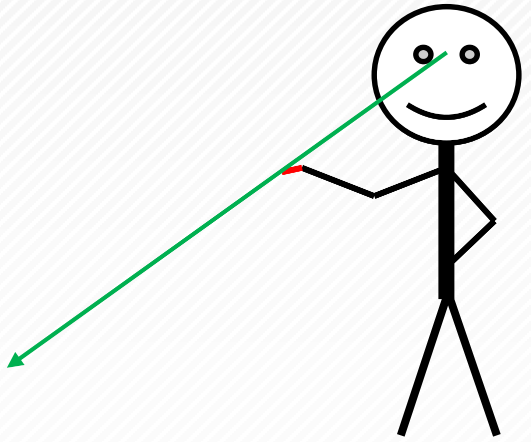


First tentative:

- Direction: Finger direction
- Position: Finger position

Results:

- Lot of errors (targeting too far)
- User point of view: Hard to understand where he is actually pointing
  - Cannot see where his finger is really pointing at. Just a rough idea.



Second tentative :

- Direction: Eyes → Tip of the finger
- Position: Eyes

Results:

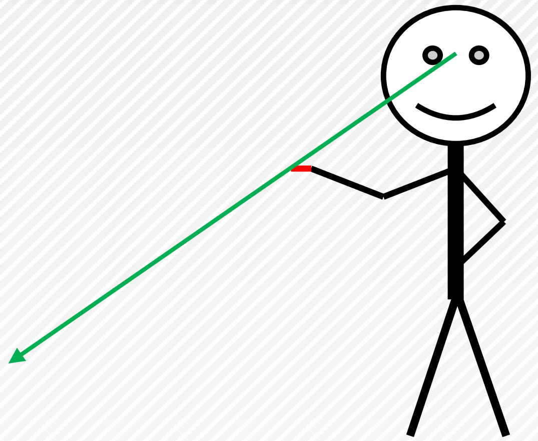
- Less errors, more accurate, but depends on the user
- User point of view: Easy to understand where they are actually pointing
  - Can see what they are targeting.

POINTING AT LOCATION, OBJECTS

POINTING METHOD

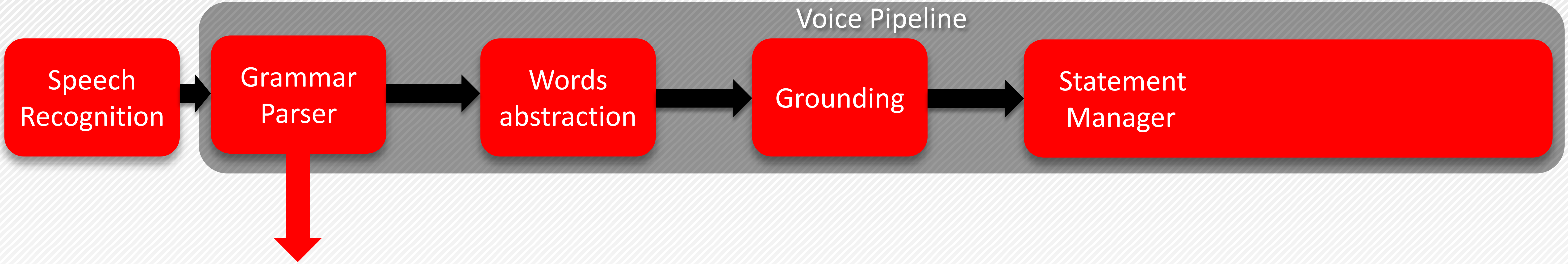
Player can point:

- With the pointing finger of their choice
  - Direction: Eyes → Tip of the finger
  - Position: Eyes
- With their eyes only



LOCATION-BASED INFORMATION

DISAMBIGUATION


Voice Pipeline

Speech Recognition

Grammar Parser

Words abstraction

Grounding

Statement Manager

- Put the apple on the left of the box.

  [Verb: Put] [Object: apple] [Location: left] [Object: box]

- Go behind the rocket.

  [Verb: Go] [Location: behind] [Object: rocket]

- Take the apple that is on the table.

  [Verb: Take] [Object: apple] [Description: that is] [Location: on] [Object: table]

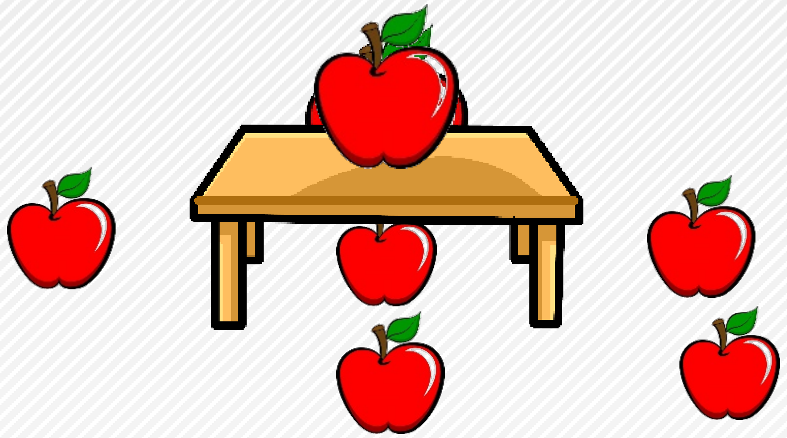
LOCATION-BASED INFORMATION

DISAMBIGUATION

- List of locations:
  - Next to / Away from
  - On / Under
  - Left / Right
  - Front / Behind

Easy

Complex

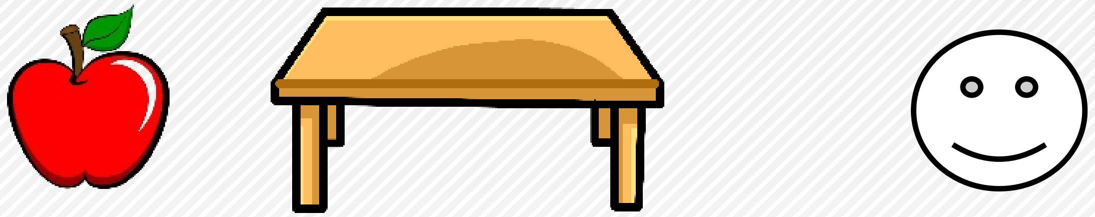


LOCATION-BASED INFORMATION

DISAMBIGUATION

Where is the apple?

- Behind the table?



→ Depends on the player location

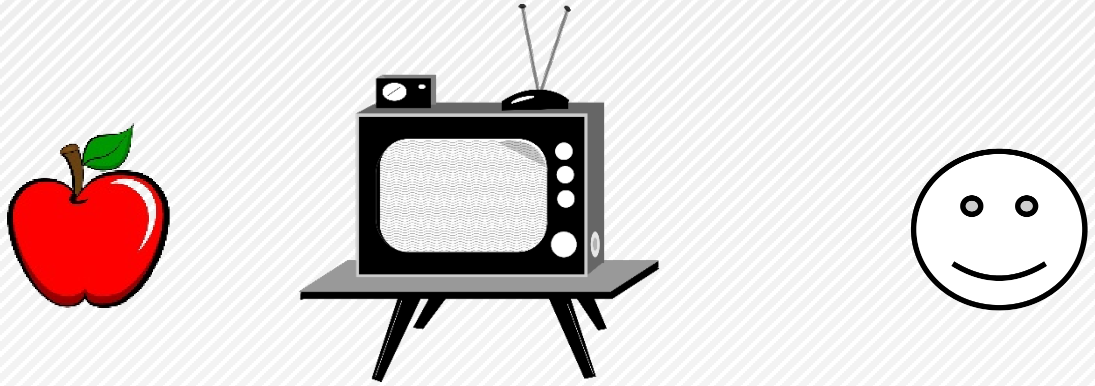
→ In this case, it is “behind”

LOCATION-BASED INFORMATION

DISAMBIGUATION

Where is the apple?

- Behind the TV ?



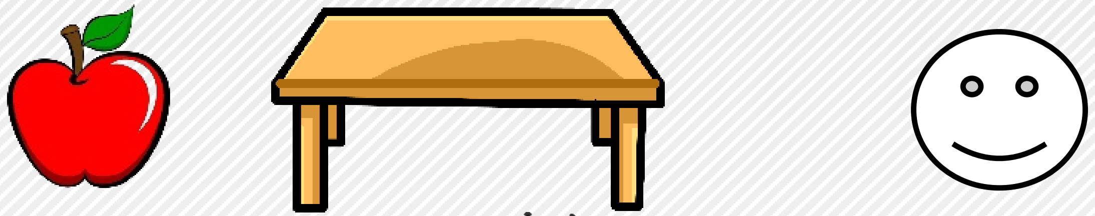
→ Depends on the object type

→ In this case, it is “on the left”

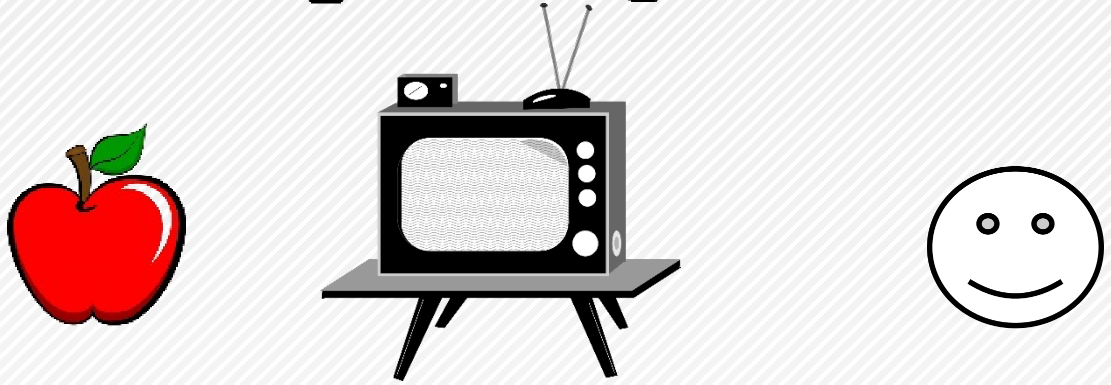
LOCATION-BASED INFORMATION

DISAMBIGUATION

Apple is behind the table



Apple is on the left of the TV



“Left / Right / Front / Behind” disambiguation depends on:
→ Object type
→ Player point of view

Solution: “Does the reference-object have an orientation ?”

	Depends on:
➡ Yes:	Object orientation
➡ No:	Player point of view

WHAT DID WE ACHIEVED SO FAR

- Bring more natural interactions:
    - Voice interactions
        - Speech recognition pipeline (faster and more direct interactions)
        - Location-based information
    - Body interactions
        - Pointing at locations while speaking

- Multi-language support for speech recognition can be achieved in a sort-of general manner.
    - The grammar parser still need to be created for each language.

WHAT CAN WE DO FROM HERE?

- Explore other solutions for failure cases where there is no very good solution yet.
- Multi-agents
- Support more kind of statements
  - Questions, Empathy…
- More interaction from the agent to the Player

Enhanced Immersivity: Using Speech Recognition for More Natural Player AI Interactions

Gautier Boeda
AI Engineer – SQUARE ENIX CO., LTD
boedagau@square-enix.com