The current situation and long-term perspectives of Hungarian language technology

Kornai András
MTA SZTAKI
BME Department of Algebra

CESAR

January 18 2013

When do we consider HLT good?

- When we don't notice it
- When it is error free
- When it becomes part of our everyday life
- When it expands our potential

When do we consider HLT good?

- When we don't notice it
- When it is error free
- When it becomes part of our everyday life
- When it expands our potential

When do we consider HLT good?

- When we don't notice it
- When it is error free
- When it becomes part of our everyday life
- When it expands our potential

When do we consider HLT good?

- When we don't notice it
- When it is error free
- When it becomes part of our everyday life
- When it expands our potential

When do we consider HLT good?

- When we don't notice it
- When it is error free
- When it becomes part of our everyday life
- When it expands our potential

Examples

- Spellchecking
  - Not exactly unnoticeable
  - But indispensable
  - Not error free
  - But who dares to argue with the red underlining
- Web search
  - Not unnoticeable
  - Completely indispensable
  - Not error free
  - But the user doesn't know this

Examples

- Spellchecking
  1. Not exactly unnoticeable
  2. But indispensable
  3. Not error free
  4. But who dares to argue with the red underlining
- Web search
  1. Not unnoticeable
  2. Completely indispensable
  3. Not error free
  4. But the user doesn't know this

Examples

- Spellchecking
  1. Not exactly unnoticeable
  2. But indispensable
  3. Not error free
  4. But who dares to argue with the red underlining
- Web search
  1. Not unnoticeable
  2. Completely indispensable
  3. Not error free
  4. But the user doesn't know this

Examples

- Spellchecking
  1. Not exactly unnoticeable
  2. But indispensable
  3. Not error free
  4. But who dares to argue with the red underlining
- Web search
  1. Not unnoticeable
  2. Completely indispensable
  3. Not error free
  4. But the user doesn't know this

Examples

- Spellchecking
  1. Not exactly unnoticeable
  2. But indispensable
  3. Not error free
  4. But who dares to argue with the red underlining
- Web search
  1. Not unnoticeable
  2. Completely indispensable
  3. Not error free
  4. But the user doesn't know this

Examples

- Spellchecking
  1. Not exactly unnoticeable
  2. But indispensable
  3. Not error free
  4. But who dares to argue with the red underlining
- Web search
  1. Not unnoticeable
  2. Completely indispensable
  3. Not error free
  4. But the user doesn't know this

Examples

- Spellchecking
  1. Not exactly unnoticeable
  2. But indispensable
  3. Not error free
  4. But who dares to argue with the red underlining
- Web search
  1. Not unnoticeable
  2. Completely indispensable
  3. Not error free
  4. But the user doesn't know this

Examples

- Spellchecking
  1. Not exactly unnoticeable
  2. But indispensable
  3. Not error free
  4. But who dares to argue with the red underlining
- Web search
  1. Not unnoticeable
  2. Completely indispensable
  3. Not error free
  4. But the user doesn't know this

Examples

- Spellchecking
  1. Not exactly unnoticeable
  2. But indispensable
  3. Not error free
  4. But who dares to argue with the red underlining
- Web search
  1. Not unnoticeable
  2. Completely indispensable
  3. Not error free
  4. But the user doesn't know this

Examples

- Spellchecking
  1. Not exactly unnoticeable
  2. But indispensable
  3. Not error free
  4. But who dares to argue with the red underlining
- Web search
  1. Not unnoticeable
  2. Completely indispensable
  3. Not error free
  4. But the user doesn't know this

Examples

- Spellchecking
  1. Not exactly unnoticeable
  2. But indispensable
  3. Not error free
  4. But who dares to argue with the red underlining
- Web search
  1. Not unnoticeable
  2. Completely indispensable
  3. Not error free
  4. But the user doesn't know this

Language technologies in development

Are very much like Machine Translation

- Not unnoticeable (but it would be if it were good)
- Not error free (but it would be if it were good)
- Not part of our everyday lives (but it would be if it were good)
- But it already expands our potential

Language technologies in development

Are very much like Machine Translation

- Not unnoticeable (but it would be if it were good)
- Not error free (but it would be if it were good)
- Not part of our everyday lives (but it would be if it were good)
- But it already expands our potential

Language technologies in development

Are very much like Machine Translation

- Not unnoticeable (but it would be if it were good)
- Not error free (but it would be if it were good)
- Not part of our everyday lives (but it would be if it were good)
- But it already expands our potential

Language technologies in development

Are very much like Machine Translation

- Not unnoticeable (but it would be if it were good)
- Not error free (but it would be if it were good)
- Not part of our everyday lives (but it would be if it were good)
- But it already expands our potential

Language technologies in development

Are very much like Machine Translation

- Not unnoticeable (but it would be if it were good)
- Not error free (but it would be if it were good)
- Not part of our everyday lives (but it would be if it were good)
- But it already expands our potential

What does it take to build good HLT?

- Largely correct theoretical model (e.g. Newtonian mechanics)
- Largely correct computational mechanism
- Largely correct/complete data
- Societal responsiveness

What does it take to build good HLT?

- Largely correct theoretical model (e.g. Newtonian mechanics)
- Largely correct computational mechanism
- Largely correct/complete data
- Societal responsiveness

What does it take to build good HLT?

- Largely correct theoretical model (e.g. Newtonian mechanics)
- Largely correct computational mechanism
- Largely correct/complete data
- Societal responsiveness

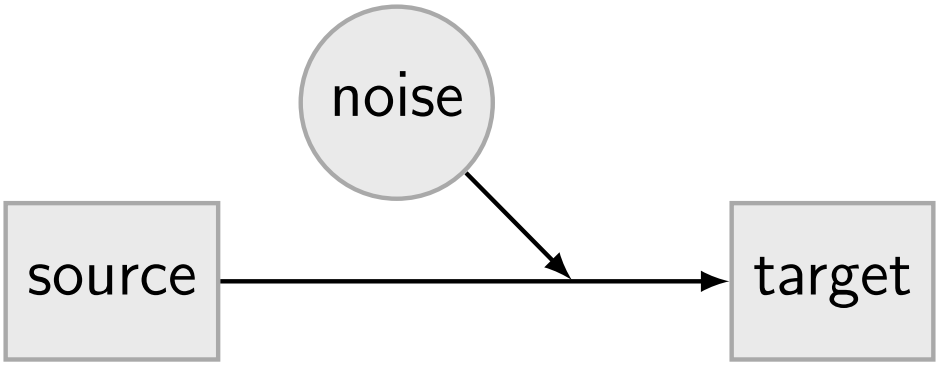
What does it take to build good HLT?

- Largely correct theoretical model (e.g. Newtonian mechanics)
- Largely correct computational mechanism
- Largely correct/complete data
- Societal responsiveness

What does it take to build good HLT?

- Largely correct theoretical model (e.g. Newtonian mechanics)
- Largely correct computational mechanism
- Largely correct/complete data
- Societal responsiveness

The classical framework (Shannon-Weaver 1948)



Changes since 1948

- The classical framework was well suited for the traditional channels of communication telephone, telex, printed matter
- These channels were not capable of processing, selecting, or transforming information – these tasks were left entirely to the sender and the receiver.
- The chips in smartphones and in the computers that make up the internet have high computing power, which provides the foundation for the channel itself to provide a part of the intelligence.

Changes since 1948

- The classical framework was well suited for the traditional channels of communication telephone, telex, printed matter
- These channels were not capable of processing, selecting, or transforming information – these tasks were left entirely to the sender and the receiver.
- The chips in smartphones and in the computers that make up the internet have high computing power, which provides the foundation for the channel itself to provide a part of the intelligence.

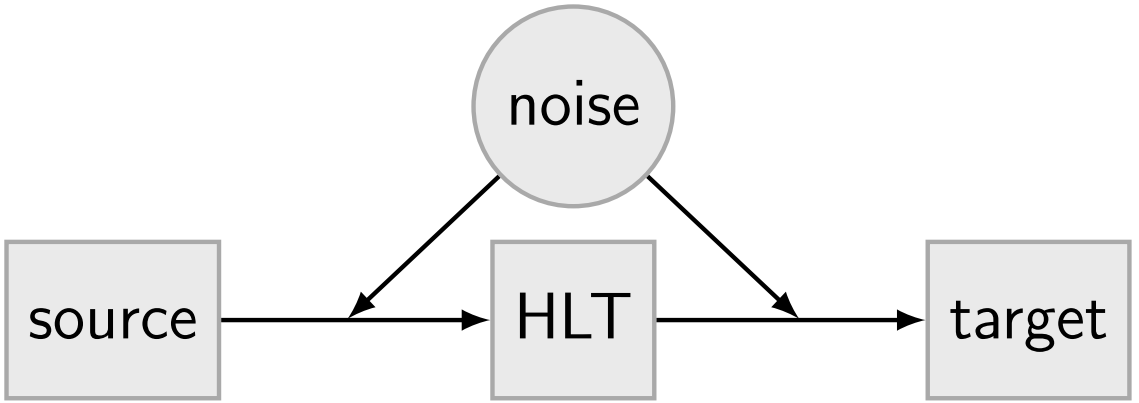
Changes since 1948

- The classical framework was well suited for the traditional channels of communication telephone, telex, printed matter
- These channels were not capable of processing, selecting, or transforming information – these tasks were left entirely to the sender and the receiver.
- The chips in smartphones and in the computers that make up the internet have high computing power, which provides the foundation for the channel itself to provide a part of the intelligence.

Changes since 1948

- The classical framework was well suited for the traditional channels of communication telephone, telex, printed matter
- These channels were not capable of processing, selecting, or transforming information – these tasks were left entirely to the sender and the receiver.
- The chips in smartphones and in the computers that make up the internet have high computing power, which provides the foundation for the channel itself to provide a part of the intelligence.

Shannon-Weaver 2.0



(HLT – Human Language Technology)

Examples

- Apple Spotlight – can find the form *went* if the query was *go*
- Browser can give cross-language pages (based e.g. on Google Translate)
- Microsoft speech translator

Examples

- Apple Spotlight – can find the form *went* if the query was *go*
- Browser can give cross-language pages (based e.g. on Google Translate)
- Microsoft speech translator

Examples

- Apple Spotlight – can find the form *went* if the query was *go*
- Browser can give cross-language pages (based e.g. on Google Translate)
- Microsoft speech translator

Examples

- Apple Spotlight – can find the form *went* if the query was *go*
- Browser can give cross-language pages (based e.g. on Google Translate)
- Microsoft speech translator

Unnoticeable components

Like the carburetor: it is very hard to build a good one, it is indispensable for a working car, but the end-user is not at all interested

- Tokenization – where words and sentences begin and end
- Paer of speech tagging – nouns, verbs, adjectives, ...
- Morphological analysis – inflection, derivation, compounding, ...
- Named entity recognition – names of people, places, ...
- Chunking – finding Noun Phrases and other surface constituents
- Functional structure – subject, object, indirect object, ...
- Anaphors, pronoun resolution, dialog management, ...
- Logical inference, world model, ontology, ...
- ...

Unnoticeable components

Like the carburetor: it is very hard to build a good one, it is indispensable for a working car, but the end-user is not at all interested

- Tokenization – where words and sentences begin and end
- Paer of speech tagging – nouns, verbs, adjectives, ...
- Morphological analysis – inflection, derivation, compounding, ...
- Named entity recognition – names of people, places, ...
- Chunking – finding Noun Phrases and other surface constituents
- Functional structure – subject, object, indirect object, ...
- Anaphors, pronoun resolution, dialog management, ...
- Logical inference, world model, ontology, ...
- ...

Unnoticeable components

Like the carburetor: it is very hard to build a good one, it is indispensable for a working car, but the end-user is not at all interested

- Tokenization – where words and sentences begin and end
- Paer of speech tagging – nouns, verbs, adjectives, ...
- Morphological analysis – inflection, derivation, compounding, ...
- Named entity recognition – names of people, places, ...
- Chunking – finding Noun Phrases and other surface constituents
- Functional structure – subject, object, indirect object, ...
- Anaphors, pronoun resolution, dialog management, ...
- Logical inference, world model, ontology, ...
- ...

Unnoticeable components

Like the carburetor: it is very hard to build a good one, it is indispensable for a working car, but the end-user is not at all interested

- Tokenization – where words and sentences begin and end
- Paer of speech tagging – nouns, verbs, adjectives, ...
- Morphological analysis – inflection, derivation, compounding, ...
- Named entity recognition – names of people, places, ...
- Chunking – finding Noun Phrases and other surface constituents
- Functional structure – subject, object, indirect object, ...
- Anaphors, pronoun resolution, dialog management, ...
- Logical inference, world model, ontology, ...
- ...

Unnoticeable components

Like the carburetor: it is very hard to build a good one, it is indispensable for a working car, but the end-user is not at all interested

- Tokenization – where words and sentences begin and end
- Paer of speech tagging – nouns, verbs, adjectives, ...
- Morphological analysis – inflection, derivation, compounding, ...
- Named entity recognition – names of people, places, ...
- Chunking – finding Noun Phrases and other surface constituents
- Functional structure – subject, object, indirect object, ...
- Anaphors, pronoun resolution, dialog management, ...
- Logical inference, world model, ontology, ...
- ...

Unnoticeable components

Like the carburetor: it is very hard to build a good one, it is indispensable for a working car, but the end-user is not at all interested

- Tokenization – where words and sentences begin and end
- Paer of speech tagging – nouns, verbs, adjectives, ...
- Morphological analysis – inflection, derivation, compounding, ...
- Named entity recognition – names of people, places, ...
- Chunking – finding Noun Phrases and other surface constituents
- Functional structure – subject, object, indirect object, ...
- Anaphors, pronoun resolution, dialog management, ...
- Logical inference, world model, ontology, ...
- ...

Unnoticeable components

Like the carburetor: it is very hard to build a good one, it is indispensable for a working car, but the end-user is not at all interested

- Tokenization – where words and sentences begin and end
- Paer of speech tagging – nouns, verbs, adjectives, ...
- Morphological analysis – inflection, derivation, compounding, ...
- Named entity recognition – names of people, places, ...
- Chunking – finding Noun Phrases and other surface constituents
- Functional structure – subject, object, indirect object, ...
- Anaphors, pronoun resolution, dialog management, ...
- Logical inference, world model, ontology, ...
- ...

Unnoticeable components

Like the carburetor: it is very hard to build a good one, it is indispensable for a working car, but the end-user is not at all interested

- Tokenization – where words and sentences begin and end
- Paer of speech tagging – nouns, verbs, adjectives, ...
- Morphological analysis – inflection, derivation, compounding, ...
- Named entity recognition – names of people, places, ...
- Chunking – finding Noun Phrases and other surface constituents
- Functional structure – subject, object, indirect object, ...
- Anaphors, pronoun resolution, dialog management, ...
- Logical inference, world model, ontology, ...
- ...

Unnoticeable components

Like the carburetor: it is very hard to build a good one, it is indispensable for a working car, but the end-user is not at all interested

- Tokenization – where words and sentences begin and end
- Paer of speech tagging – nouns, verbs, adjectives, ...
- Morphological analysis – inflection, derivation, compounding, ...
- Named entity recognition – names of people, places, ...
- Chunking – finding Noun Phrases and other surface constituents
- Functional structure – subject, object, indirect object, ...
- Anaphors, pronoun resolution, dialog management, ...
- Logical inference, world model, ontology, ...
- ...

Unnoticeable components

Like the carburetor: it is very hard to build a good one, it is indispensable for a working car, but the end-user is not at all interested

- Tokenization – where words and sentences begin and end
- Paer of speech tagging – nouns, verbs, adjectives, ...
- Morphological analysis – inflection, derivation, compounding, ...
- Named entity recognition – names of people, places, ...
- Chunking – finding Noun Phrases and other surface constituents
- Functional structure – subject, object, indirect object, ...
- Anaphors, pronoun resolution, dialog management, ...
- Logical inference, world model, ontology, ...
- ...

Hungarian language technology based on English models?

- Feasible for speech, but the English model can't take advantage of the major regularities of Hungarian speech such as vowel harmony
- Not feasible at the word level, since the English model can't cope with the complex Hungarian patterns of suffixation
- Feasible for analyzing NPs, since the English model must be prepared for about the same complexity as the Hungarian model (fixed word order within NPs)
- Not feasible at the sentence level, since the English model is based on a condition not met by Hungarian (fixed word order)

Hungarian language technology based on English models?

- Feasible for speech, but the English model can't take advantage of the major regularities of Hungarian speech such as vowel harmony
- Not feasible at the word level, since the English model can't cope with the complex Hungarian patterns of suffixation
- Feasible for analyzing NPs, since the English model must be prepared for about the same complexity as the Hungarian model (fixed word order within NPs)
- Not feasible at the sentence level, since the English model is based on a condition not met by Hungarian (fixed word order)

Hungarian language technology based on English models?

- Feasible for speech, but the English model can't take advantage of the major regularities of Hungarian speech such as vowel harmony
- Not feasible at the word level, since the English model can't cope with the complex Hungarian patterns of suffixation
- Feasible for analyzing NPs, since the English model must be prepared for about the same complexity as the Hungarian model (fixed word order within NPs)
- Not feasible at the sentence level, since the English model is based on a condition not met by Hungarian (fixed word order)

Hungarian language technology based on English models?

- Feasible for speech, but the English model can't take advantage of the major regularities of Hungarian speech such as vowel harmony
- Not feasible at the word level, since the English model can't cope with the complex Hungarian patterns of suffixation
- Feasible for analyzing NPs, since the English model must be prepared for about the same complexity as the Hungarian model (fixed word order within NPs)
- Not feasible at the sentence level, since the English model is based on a condition not met by Hungarian (fixed word order)

What does Hungarian HLT require?

- Largely correct theoretical model
- Largely correct computational mechanism
- Largely correct/complete data
- Societal responsiveness

What does Hungarian HLT require?

- Largely correct theoretical model
- Largely correct computational mechanism
- Largely correct/complete data
- Societal responsiveness

What does Hungarian HLT require?

- Largely correct theoretical model
- Largely correct computational mechanism
- Largely correct/complete data
- Societal responsiveness

What does Hungarian HLT require?

- Largely correct theoretical model
- Largely correct computational mechanism
- Largely correct/complete data
- Societal responsiveness

What does Hungarian HLT require?

- Largely correct theoretical model
- Largely correct computational mechanism
- Largely correct/complete data
- Societal responsiveness

What is the state of teh art?

- Largely correct theoretical model – still requires a lot of work
- Largely correct computational mechanism – we are doing well
- Largely correct/complete data – we are not doing too well
- Societal responsiveness – not without problems

What is the state of teh art?

- Largely correct theoretical model – still requires a lot of work
- Largely correct computational mechanism – we are doing well
- Largely correct/complete data – we are not doing too well
- Societal responsiveness – not without problems

What is the state of teh art?

- Largely correct theoretical model – still requires a lot of work
- Largely correct computational mechanism – we are doing well
- Largely correct/complete data – we are not doing too well
- Societal responsiveness – not without problems

What is the state of teh art?

- Largely correct theoretical model – still requires a lot of work
- Largely correct computational mechanism – we are doing well
- Largely correct/complete data – we are not doing too well
- Societal responsiveness – not without problems

What is the state of teh art?

- Largely correct theoretical model – still requires a lot of work
- Largely correct computational mechanism – we are doing well
- Largely correct/complete data – we are not doing too well
- Societal responsiveness – not without problems

Major researh centers

- Hungarian Academy of Sciences (Institute of Linguistics, Computer and Automation Research Institute, ...)
- Universities (BUTE, Debrecen, PPCU, Szeged,...)
- Hungarian industry (Aitia, ALL, Kilgray, Morphologic,...)
- Hongarian products by multinationals (Apple, Google, IBM, Microsoft, Nuance, Xerox,...)

Major researh centers

- Hungarian Academy of Sciences (Institute of Linguistics, Computer and Automation Research Institute, ...)
- Universities (BUTE, Debrecen, PPCU, Szeged,...)
- Hungarian industry (Aitia, ALL, Kilgray, Morphologic,...)
- Hongarian products by multinationals (Apple, Google, IBM, Microsoft, Nuance, Xerox,...)

Major researh centers

- Hungarian Academy of Sciences (Institute of Linguistics, Computer and Automation Research Institute, ...)
- Universities (BUTE, Debrecen, PPCU, Szeged,...)
- Hungarian industry (Aitia, ALL, Kilgray, Morphologic,...)
- Hongarian products by multinationals (Apple, Google, IBM, Microsoft, Nuance, Xerox,...)

Major researh centers

- Hungarian Academy of Sciences (Institute of Linguistics, Computer and Automation Research Institute, ...)
- Universities (BUTE, Debrecen, PPCU, Szeged,...)
- Hungarian industry (Aitia, ALL, Kilgray, Morphologic,...)
- Hongarian products by multinationals (Apple, Google, IBM, Microsoft, Nuance, Xerox,...)

Major researh centers

- Hungarian Academy of Sciences (Institute of Linguistics, Computer and Automation Research Institute, ...)
- Universities (BUTE, Debrecen, PPCU, Szeged,...)
- Hungarian industry (Aitia, ALL, Kilgray, Morphologic,...)
- Hongarian products by multinationals (Apple, Google, IBM, Microsoft, Nuance, Xerox,...)

Thank you for your attention!

Societal responsiveness

An article from a leading Hungarian technology blog about SZTAKI's language-driven railroad ticket clerk
[Click through the above line only if you read Hungarian]

Societal responsiveness II

Erika | 2012. 09. 28. 21:06 2

Tehát az eddig is közutálatnak örvendő hangosmenü géphangot átveszi a szoftver... ez biztos sokkal közkedveltebb és népszerübb lesz, kérdés csupán az, hogy mitől???

caramell | 2012. 09. 28. 20:17 1

"Kornai András szerint a megoldásuk nem csupán a MÁV-pénztáros szerepét tudná átvenni. A kutató szerint az összes olyan emberi munkát ki lehetne váltani szoftverrel, ahová csak azért ültetnek embert, mert az képes mondatokat értelmezni. Még akár a telefonos ügyfélszolgálatok közutálatnak örvendő hangosmenüjének helyét is átveheti egy olyan szoftver, aminek el kell mondani, hogy milyen panasszal telefonáltak, és rögtön kapcsolja a megfelelő ügyintézős."

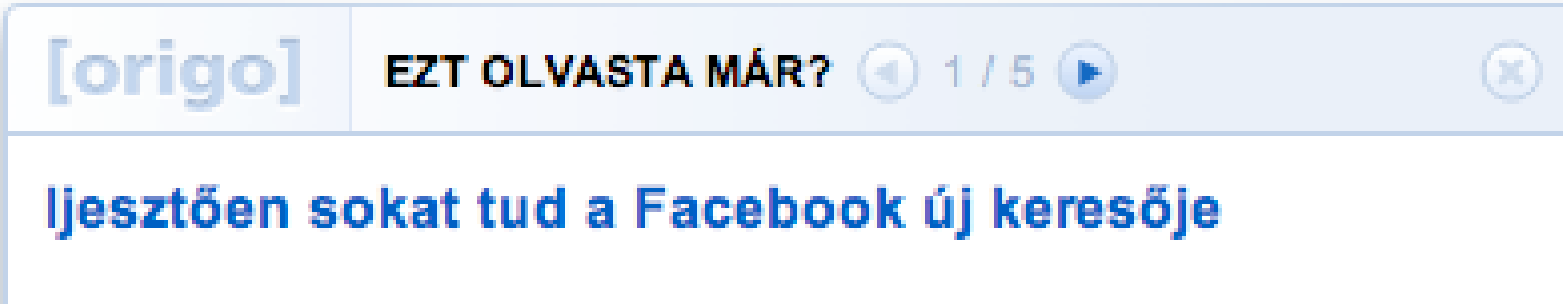
HATÁROZOTTAN ELLENE VAGYOK,IDEJE LENNE HATÁRT SZABNI AZ EMBEREK FÖLDÖNFUTÓVÁ TÉTELÉNEK!NEM ERRE KELLENE HASZNÁLNI AZ INFORMATIKÁT MOST ILYEN EGYÉBKÉNT IS NEHÉZ GAZDASÁGI VÁLSÁGBAN!

EZZEL NEM MUNKAHELYET VÉDENEK!,HANEM MEGSEMMISITENEK.

CSAK A PROFIT SZÁMIT!? AZ EMBER,A CSALÁDOK FÖLDÖNFUTÓVÁ TÉTELE NEM?
EZZEL CSAK AZ 1-2 EGYÉBKÉNT IS MILLIÁRDOS INFORMATIKUST TESZIK MÉG GAZDAGABBÁ,MIKÖZBEN AZ EMBEREK EZREIT,VAGY MILLIÓIT MUNKANÉLKÜLIVÉ TESZNEK.

[Reader comment on the article, accusing ‘billionaire informatics experts’ of destroying jobs]

Societal responsiveness III



[Recommendation from the same tech blog – Facebook's new search 'knows a frightening amount']