Tilraun til upplýsingaútdráttar

Hulda Óladóttir

Yfirlit

- Upplýsingaútdráttur og vélrænar þýðingar
- Sérsvið: Veður og náttúruhamfarir
- Upplýsingarammar
- Auðvelt að flytja á milli sérsviða
  - Málheild
  - Merkingarflokkar og hlutverk
  - Sáðorð innan merkingarflokka

Grunnhugtök

- Upplýsingaútdráttur finnur:
  - Heiti hluta
  - Vensl þeirra á milli
  - Hlutverk þeirra í atburðum
- Millivegur lykilorðaleitar og textaskilnings

Aðföng og hjálpargögn 1

- Málheild: 412 skjöl úr Morgunblaðinu (330 þjálfun, 82, prófanir)
- IceNLP til þáttunar
- Sex merkingarflokkar
  - Veður [atburður]: veður, rigning, snjór, snjókoma, frost, sól
  - Hamfarir [atburður]: eldgos, jarðskjálfti, flóð, snjóflóð, skjálfti
  - Tímasetning [tímas.]: morgunn, þriðjudagur, helgi, klukka, kvöld
  - Staðsetning [staðs.]: vesturland, austurland, reykjavík, akureyri, suðurland
  - Hitastig [ástand]: hiti, hitastig, gráða, lofthiti, frostmark
  - Manneskja [manneskja]: björgunarsveit, lögregla, landhelgisgæsla, slökkvilið, leitarhundur

Aðföng og hjálpargögn 2

- Leiðsagnarreglur

Ef nafnliðurinn er frumlag

<frl> germyndarsögn	<hitinn> lækkaði
<frl> þolmyndarsögn	<hitinn> var lækkaður
<frl> sögn sagnfylling	<skýið> var óveðurssky

Ef nafnliðurinn er beint andlag

germyndarsögn <b.andl.>	sendi <gjöfina>
þolmyndarsögn <b.andl.>	send <manninum>

Ef nafnliðurinn er óbeint andlag

germyndarsögn <ób.andl.>	sendi <manninum> gjöfina

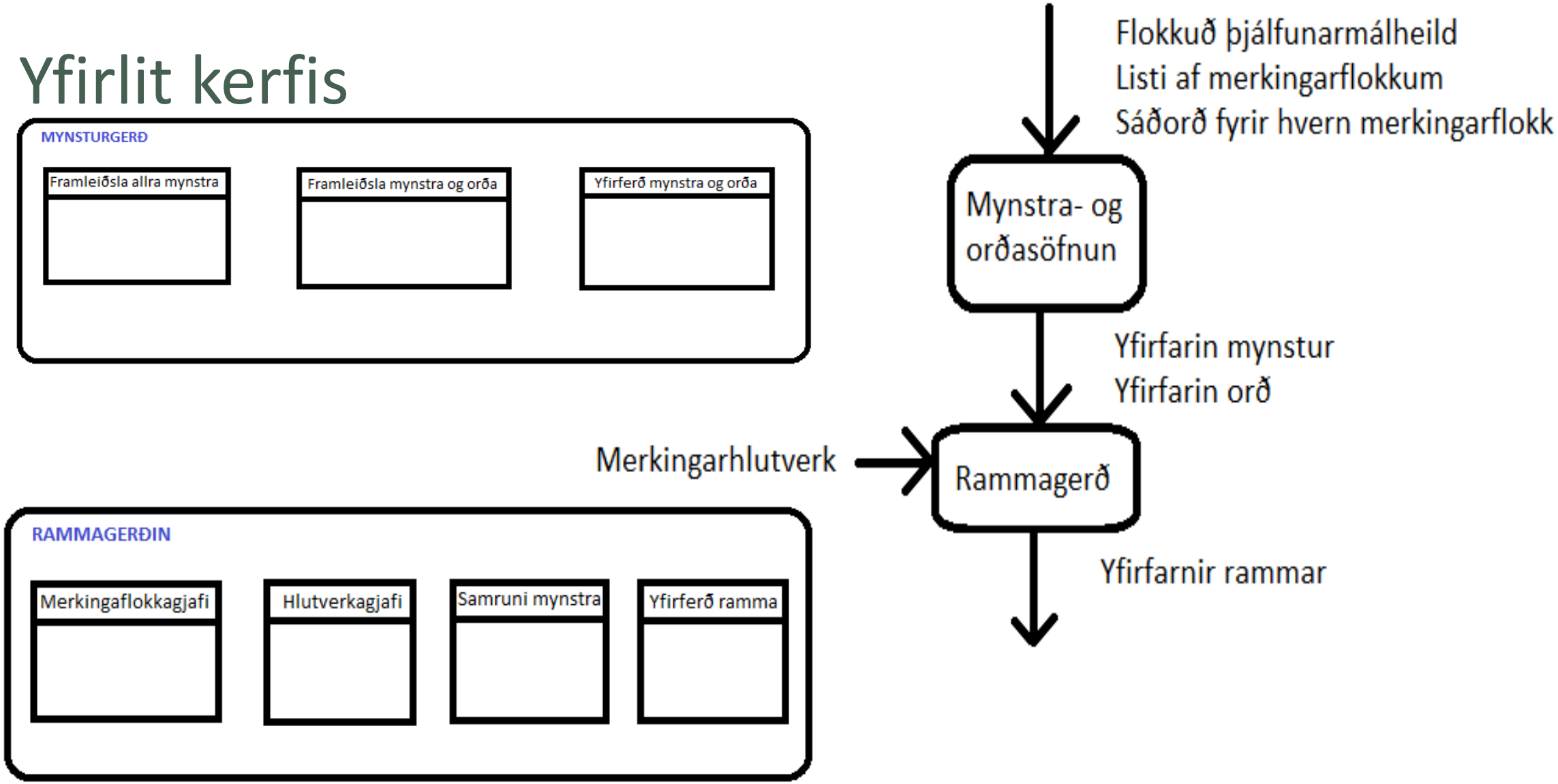
Ef nafnliðurinn er sagnfylling

nafnorð sögn <sagnfylling>	hinn látni var <fórnarlamb>

Ef nafnliðurinn er innan forsetningarliðar

no. fs. <NL>	skjálfti við <staðsetning>
germyndarsögn fs. <NL>	lækkaði í <frostmark>
þolmyndarsögn fs. <NL>	var lækkaður í <frostmark>

Yfirlit kerfis



Söfnun mynstra

- Mynstrin byggja á nafnliðum
- Leiðsagnarreglur finna virkja

- Dæmi:
  - Vísindamaðurinn mælir hitastig sjávar.
  - Veðrið á Akureyri er milt.

Ef nafnliðurinn er frumlag

<frl> germyndarsögn	<hitinn> lækkaði
<frl> þolmyndarsögn	<hitinn> var lækkaður
<frl> sögn sagnfylling	<skýið> var óveðurssky

Ef nafnliðurinn er beint andlag

germyndarsögn <b.andl.>	sendi <gjöfina>
þolmyndarsögn <b.andl.>	send <manninum>

Ef nafnliðurinn er óbeint andlag

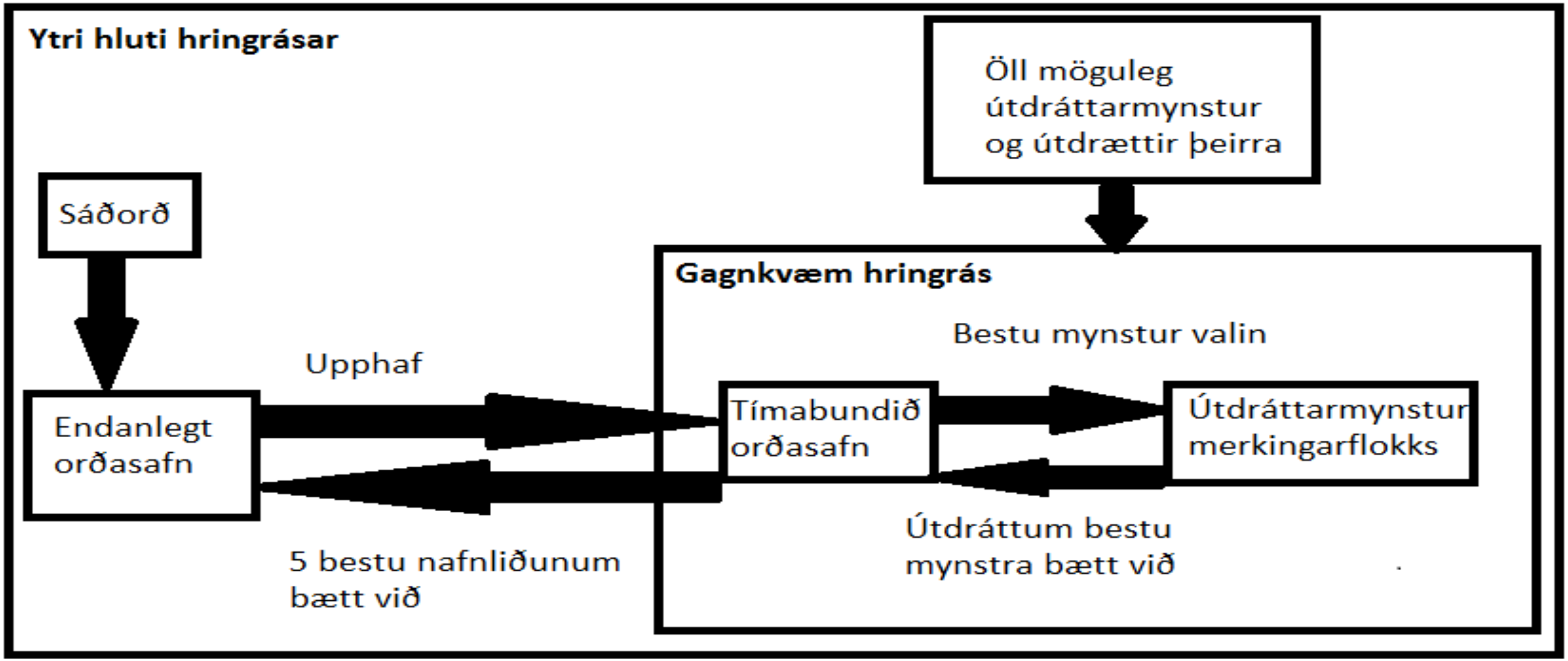
germyndarsögn <ób.andl.>	sendi <manninum> gjöfina

Ef nafnliðurinn er sagnfylling

nafnorð sögn <sagnfylling>	hinn látni var <fórnarlamb>

Ef nafnliðurinn er innan forsetningarliðar

no. fs. <NL>	skjálfti við <staðsetning>
germyndarsögn fs. <NL>	lækkaði í <frostmark>
þolmyndarsögn fs. <NL>	var lækkaður í <frostmark>



- $einkunn\ mynsturs\ i = R_i * \log_2(F_i+1)$
- $einkunn\ nafnliðar\ i = \sum_{k=1}^{N_i} 1 + (0{,}01 * einkunn(mynstur\ k))$

Endanlegt safn orða og mynstra

- Öll mynstur voru 4.290 talsins
- Kerfið fann 670 mikilvæg mynstur
- Eftir yfirferð stóðu 574 mynstur eftir
- Endanlega orðasafnið innihélt 581 orð
  - Veður: 124
  - Hamfarir: 104
  - Staðsetning: 152
  - Tímasetning: 110
  - Hitastig: 14
  - Manneskja: 77

Merkingarlýsingu bætt við mynstur

- Hvaða merkingarflokkum tengist mynstrið?
- Sterk vensl geta orðið merkingarhömlur

- Dæmi um mynstur og merkingarhömlur:
- noun|acc|veður|noun pp|á staðsetning, tímasetning
- Veður á Akureyri, veður á morgun

Mynstur sameinuð í upplýsingaramma

- Rammar tengja upplýsingar í heildstæða mynd
- Mynstur með sama virkja geta sameinast ef setningagerð leyfir

<frumlag> sprengja (gm.) gerandi, hlutv. hryðjuverkamaður
sprengja (gm.) <beint andlag> skotmark, hlutv. bygging eða farartæki
sprengja (gm.) í <nafnliður> staðsetning
Virki: sprengja (gm.)

gerandi	frumlag	hryðjuverkamaður
skotmark	beint andlag	bygging, farartæki
staðsetning	FL(í)	staðsetning

Hryðjuverkamaðurinn sprengdi sendiráðið í Lundúnum í gær.
<Hryðjuverkamaðurinn> *sprengdi* <sendiráðið> <í Lundúnum> í gær.

- Með sameiningu fengust 200 rammar

Árangursmat – prófunargögn

- 82 skjöl til prófunar
- Svaralykill útbúinn

Réttur útdráttur – Réttur útdráttur er í réttu hólfi.
Hlutréttur útdráttur – Að hluta til réttur útdráttur í hólfi.
Rangur útdráttur – Rangur útdráttur í hólfi.
Útdrátt vantar (e. *missing*) – Enginn útdráttur er í hólfi.
Falskur útdráttur (e. *spurious*) – Útdráttur þar sem ekkert á að vera.

$$nákvæmni = \frac{réttir + tvítekningar + hlutréttir}{réttir + tvítekningar + hlutréttir + falskir + rangir}$$

$$heimt = \frac{réttir}{réttir + vantar}$$

$$F\ mæling = \frac{2 * nákvæmni * heimt}{nákvæmni + heimt}$$

Árangursmat - Söfnun útdrátta

- Fyrir rammana í heild sinni
  - Nákvæmni = 23,4%
  - Heimt = 46,5%
  - F-mæling = 31,2%
- Fyrir útdrætti réttra ramma
  - Nákvæmni = 75,2%
  - Heimt = 40%
  - F-mæling = 52,2%
- Sambærilegt kerfi fyrir ensku
  - Nákvæmni = 36%
  - Heimt = 58%
  - F-mæling = 44%
- Fyrir útdrætti allra ramma
  - Nákvæmni = 17,6%
  - Heimt = 20,1%
  - F-mæling = 18,8%

Endurbætur og prófanir á nýju sérsviði

- Sérsvið hryðjuverka prófað
- 118 fréttir til þjálfunar, fjórar til prófana
- Sömu merkingarflokkar og hlutverk
  - Gerendur [manneskja]: hryðjuverkamaður, árásarmaður, sprengjumaður, ...
  - Staðsetning [staðs.]: París, Brussel, Belgía, Frakkland, Þýskaland
  - Fórnarlömb [manneskja]: Fórnarlamb, látinn, særður, myrtur, ...
  - Vopn [verkfæri]: Sprengja, byssa, vopn, bílsprengja, hnífur
  - SkotmarkA [bygging]: Bygging, sendiráð, heimili, Louvre, Bataclan, ...
  - SkotmarkB [farartæki]: Vörubíll, bíll, bifreið, farartæki, mótorhjól

Niðurstöður

- 198 mynstur, 130 eftir yfirferð
- 1.579 orð, 402 eftir yfirferð

- Fyrir rammana í heild sinni
  - Nákvæmni = 48,3%
  - Heimt = 68,2%
  - F-mæling = 56,5%

- Fyrir útdrætti réttra ramma
  - Nákvæmni = 90,6%
  - Heimt = 54,3%
  - F-mæling = 67,9%
- Fyrir útdrætti allra ramma
  - Nákvæmni = 43,3%
  - Heimt = 32%
  - F-mæling = 36,8%

Nýting og næstu skref

- Safna upplýsingum úr íslenskum textum
- Þróa áfram og greina atburðarás og meginþemu í fréttum
- Þróa þýðingahlutann áfram
- Safna orðum innan merkingarsviðs fyrir atriðisorðaskrá
- Bæta kerfið; prófa fullþáttun

https://github.com/Holado/IE

- Phillips, W. og Riloff, E. (2007). Exploiting role-identifying nouns and expressions for information extraction. Í G. Angelova, K. Bontcheva, R. Mitkov, N. Nicolov og N. Nikolov (ritstj.), *Proceedings of the International Conference on Recent Advances in Natural Language Processing* (bls. 468-473). Borovets, Búlgaríu.
- Riloff, E. (1993). Automatically constructing a dictionary for information extraction tasks. Í *Proceedings of the eleventh national conference on Artificial intelligence* (bls. 811-816), Washington, D.C.: AAAI Press.
- Riloff, E. (1996a). Automatically generating extraction patterns from untagged text. Í *Proceedings of the thirteenth national conference on Artificial intelligence* (bls. 1044-1049). Portland, Oregon: AAAI Press.
- Riloff, E. (1996b). An empirical study of automated dictionary construction for information extraction in three domains. *Artificial intelligence, 85*(1), 101-134. doi:http://dx.doi.org/10.1016/0004-3702(95)00123-9
- Riloff, E. og Shepherd, J. (1997). A corpus-based approach for building semantic lexicons. Í *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing* (bls. 117-124). Sótt af http://www.aclweb.org/anthology/W/W97/W97-0313.pdf
- Riloff, E. og Schmelzenbach, M. (1998). An empirical approach to conceptual case frame acquisition. Í *Proceedings of the Sixth Workshop on Very Large Corpora* (bls. 49-56). Sótt af http://www.aclweb.org/anthology/W98-1106
- Riloff, E. og Jones, R. (1999). Learning dictionaries for information extraction by multi-level bootstrapping. Í *Proceedings of the Sixteenth National Conference on Artificial Intelligence* (bls. 474-479). Sótt af http://www.aaai.org/Papers/AAAI/1999/AAAI99-068.pdf
- Thelen, M. og Riloff, E. (2002). A bootstrapping method for learning semantic lexicons using extraction pattern contexts. Í *Proceedings of the ACL-02 conference on Empirical methods in natural language processing* (bls. 214-221). doi:http://dx.doi.org/10.3115/1118693.1118721