SOFTWIN

SpeD 2017

July 6-9, 2017 – Bucharest, Romania

A Rule-based Approach to Generating Large Phonetic Databases for Romanian Results of the AFLR Project *

Ștefan - Stelian Diaconescu
Monica - Mihaela Rizea
Mihaela Ionescu, Andrei Mincă, Monica Rădulescu

Research & Development Department
SOFTWIN

* This work has been carried out under the Partnerships Program in Priority Areas - PNII, developed with the support of MEN - UEFISCDI, project no. PN-II-PT-PCCA-2013-4-1451, contract no. 332/2014.

OUTLINE

- Objectives of the AFLR project
- Instruments
- Evolution of the RO Phonetic Database
- Challenges and Achievements
- Conclusions and Future Aims
- Q&A

Objectives of the AFLR* Project

- Complex phonetic knowledge bases, including rules for automatic phonetic transcriptions/syllabifications
- Dictionaries that record phonemic transcriptions for Romanian words, and syllable divisions/stress patterns using an internationally-approved standard
- Automatic Speech Recognition System based on acoustic elements as well as on linguistic knowledge

* AFLR – Analiza Fonetică a Limbii Române: Studiu şi Aplicaţii Informatice
(“Romanian Language Phonetic Analysis: Study and Applications”)

Instruments

- Abstract metalanguage (GRAALAN) especially designed for describing linguistic knowledge
- Software tools:
  - GRAALAN Compiler
  - LKT (Lexicon Knowledge Tool)
  - MKT (Morphological Knowledge Tool)
  - BDKT (Bilingual Dictionary Knowledge Tool)
  - LINK

The Process of Creating an LKB

Instruments

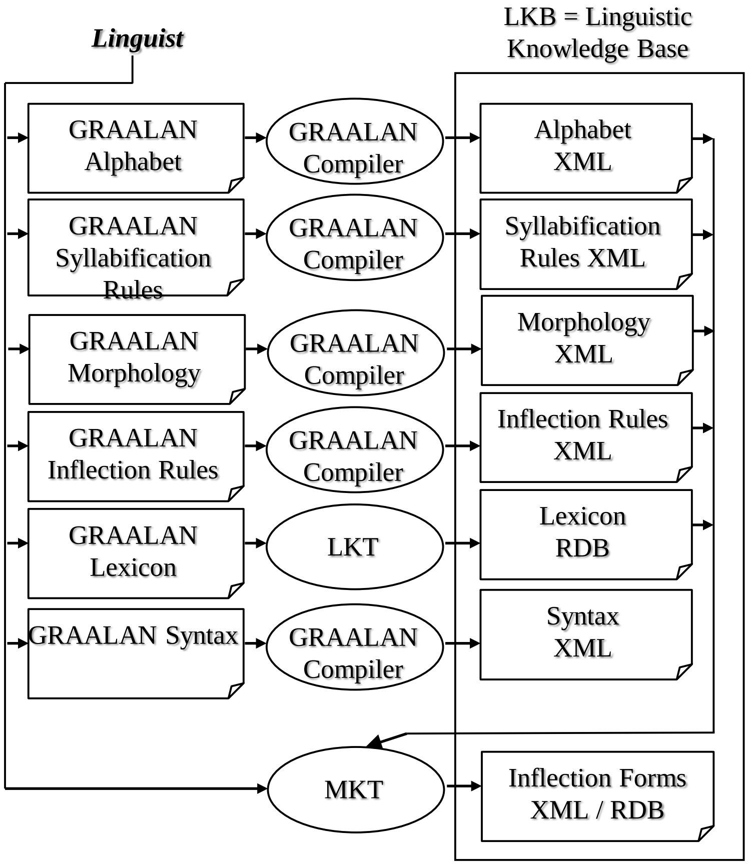


FIG. 1 Example of a GRAALAN System Operation Flow

- The linguist describes the following sections corresponding to a natural language, using GRAALAN:
  - Alphabet
  - Syllabification Rules
  - Morphology
  - Inflection Rules
  - Lexicon
  - Syntax
- The GRAALAN descriptions are compiled and converted into XML format
- The linguist creates (mostly in RDB) the lexicon of the natural language using the Lexicon Knowledge Tool (LKT)
- The paradigms (synthetic & analytical inflection forms) of all the lexicon lemmas are generated using the Morphological Knowledge Tool (MKT) (which receives as input the Alphabet, the Morphological Configurator, and the Inflection Rules in XML format)
- The paradigms are generated by applying the inflection rules to the lexicon lemmas

Instruments

- ***GRAALAN Compiler***
  - converts the linguistic descriptions from GRAALAN into XML format
- ***LKT (Lexicon Knowledge Tool)***
  - facilitates the development of the lexicon
- ***MKT (Morphological Knowledge Tool)***
  - applies the inflection rules described by the linguist to the lexicon lemmas and thus obtains all the inflection forms (in normal and phonetic alphabet)
- ***BDKT (Bilingual Dictionary Knowledge Tool)***
  - creates the correspondences between two natural languages
- ***LINK***
  - verifies the consistency of the linguistic knowledge bases

Evolution of the RO Phonetic Database in the AFLR project

- **Starting point**
- 76,837 lemmas (with phonetic transcriptions) and 116,074 meanings
- 12,119,349 inflected forms (i.e., 861,718 synthetic forms and 11,257,631 analytical forms) with phonetic transcriptions
- 13,207 multiword expressions (MWEs) (with phonetic transcriptions) and 14,052 meanings
- 90,393 semantic relations between meanings

Evolution of the RO Phonetic Database in the AFLR project

- **Final point**
- 100,708 lemmas (with phonetic transcriptions) and 142,299 meanings (- i.e., an increase of about 30% related to the n°. of lexicon lemmas)
- 13,219 MWEs (with phonetic transcriptions) and 14,066 meanings
- 184,075 semantic relations between meanings
- The inflected forms (along with their phonetic transcriptions) for the newly introduced lexicon lemmas are currently in process of being generated

Evolution of the RO Phonetic Database in the AFLR project

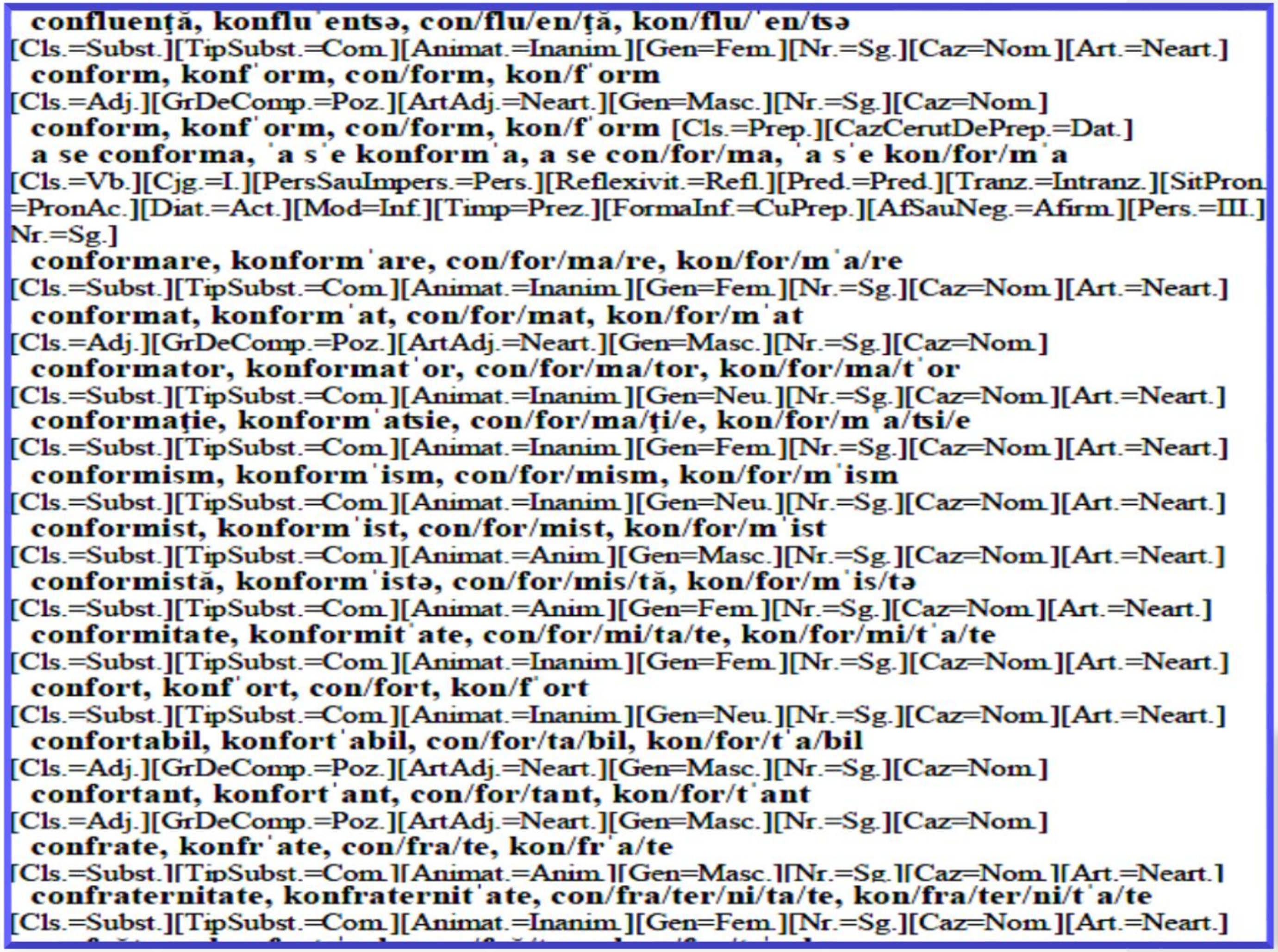
Dictionaries written in collaboration with “Iorgu Iordan – Al. Rosetti” Institute of Linguistics of the Romanian Academy (ILIR):

- The Morphological and Phonetic Dictionary
- The Phonetic Dictionary of Syllables
- The Rhyming Dictionary

The Morphological and Phonetic Dictionary – Structure of an Entry

- The word in Normal Alphabet
- The word in Phonetic Alphabet
- The word syllabification in Normal Alphabet
- The word syllabification in Phonetic Alphabet
- The morphological characterization of the word, as a sequence of morphological categories and morphological category values

The Morphological and Phonetic Dictionary



```
confluenţă, konfluˈentsə, con/flu/en/ţă, kon/flu/ˈen/tsə
[Cls.=Subst.][TipSubst.=Com.][Animat.=Inanim.][Gen=Fem.][Nr.=Sg.][Caz=Nom.][Art.=Neart.]
conform, konfˈorm, con/form, kon/fˈorm
[Cls.=Adj.][GrDeComp.=Poz.][ArtAdj.=Neart.][Gen=Masc.][Nr.=Sg.][Caz=Nom.]
conform, konfˈorm, con/form, kon/fˈorm [Cls.=Prep.][CazCerutDePrep.=Dat.]
a se conforma, ˈa sˈe konformˈa, a se con/for/ma, ˈa sˈe kon/for/mˈa
[Cls.=Vb.][Cjg.=I.][PersSauImpers.=Pers.][Reflexivit.=Refl.][Pred.=Pred.][Tranz.=Intranz.][SitPron.
=PronAc.][Diat.=Act.][Mod=Inf.][Timp=Prez.][FormaInf.=CuPrep.][AfSauNeg.=Afirm.][Pers.=III.]
Nr.=Sg.]
conformare, konformˈare, con/for/ma/re, kon/for/mˈa/re
[Cls.=Subst.][TipSubst.=Com.][Animat.=Inanim.][Gen=Fem.][Nr.=Sg.][Caz=Nom.][Art.=Neart.]
conformat, konformˈat, con/for/mat, kon/for/mˈat
[Cls.=Adj.][GrDeComp.=Poz.][ArtAdj.=Neart.][Gen=Masc.][Nr.=Sg.][Caz=Nom.]
conformator, konformatˈor, con/for/ma/tor, kon/for/ma/tˈor
[Cls.=Subst.][TipSubst.=Com.][Animat.=Inanim.][Gen=Neu.][Nr.=Sg.][Caz=Nom.][Art.=Neart.]
conformaţie, konformˈatsie, con/for/ma/ţi/e, kon/for/mˈa/tsi/e
[Cls.=Subst.][TipSubst.=Com.][Animat.=Inanim.][Gen=Fem.][Nr.=Sg.][Caz=Nom.][Art.=Neart.]
conformism, konformˈism, con/for/mism, kon/for/mˈism
[Cls.=Subst.][TipSubst.=Com.][Animat.=Inanim.][Gen=Neu.][Nr.=Sg.][Caz=Nom.][Art.=Neart.]
conformist, konformˈist, con/for/mist, kon/for/mˈist
[Cls.=Subst.][TipSubst.=Com.][Animat.=Anim.][Gen=Masc.][Nr.=Sg.][Caz=Nom.][Art.=Neart.]
conformistă, konformˈistə, con/for/mis/tă, kon/for/mˈis/tə
[Cls.=Subst.][TipSubst.=Com.][Animat.=Anim.][Gen=Fem.][Nr.=Sg.][Caz=Nom.][Art.=Neart.]
conformitate, konformitˈate, con/for/mi/ta/te, kon/for/mi/tˈa/te
[Cls.=Subst.][TipSubst.=Com.][Animat.=Inanim.][Gen=Fem.][Nr.=Sg.][Caz=Nom.][Art.=Neart.]
confort, konfˈort, con/fort, kon/fˈort
[Cls.=Subst.][TipSubst.=Com.][Animat.=Inanim.][Gen=Neu.][Nr.=Sg.][Caz=Nom.][Art.=Neart.]
confortabil, konfortˈabil, con/for/ta/bil, kon/for/tˈa/bil
[Cls.=Adj.][GrDeComp.=Poz.][ArtAdj.=Neart.][Gen=Masc.][Nr.=Sg.][Caz=Nom.]
confortant, konfortˈant, con/for/tant, kon/for/tˈant
[Cls.=Adj.][GrDeComp.=Poz.][ArtAdj.=Neart.][Gen=Masc.][Nr.=Sg.][Caz=Nom.]
confrate, konfrˈate, con/fra/te, kon/frˈa/te
[Cls.=Subst.][TipSubst.=Com.][Animat.=Anim.][Gen=Masc.][Nr.=Sg.][Caz=Nom.][Art.=Neart.]
confraternitate, konfraternitˈate, con/fra/ter/ni/ta/te, kon/fra/ter/ni/tˈa/te
[Cls.=Subst.][TipSubst.=Com.][Animat.=Inanim.][Gen=Fem.][Nr.=Sg.][Caz=Nom.][Art.=Neart.]
```

The resulting phonetic dictionaries generally make use of the principle of **broad phonetic transcription** (also named phonemic transcription).

Challenges and Achievements

- The process of enhancing the phonetic database, during AFLR, implied selecting a list of candidate lexicon entries from various digital corpora, digital dictionaries, and standardized infrastructural linguistic knowledge bases, which resulted in a list of **24,719** candidates

  - The majority of these potential entries, i.e. **21,028** words, implied neologisms (mostly not adapted to the Romanian spelling/morphology – representing common language, but also specialized vocabularies) that usually do not have pronunciation suggestions in case they are registered in the existing Romanian dictionaries – **phonetic transcriptions, syllabifications, and stress patterns have been provided for all these cases**

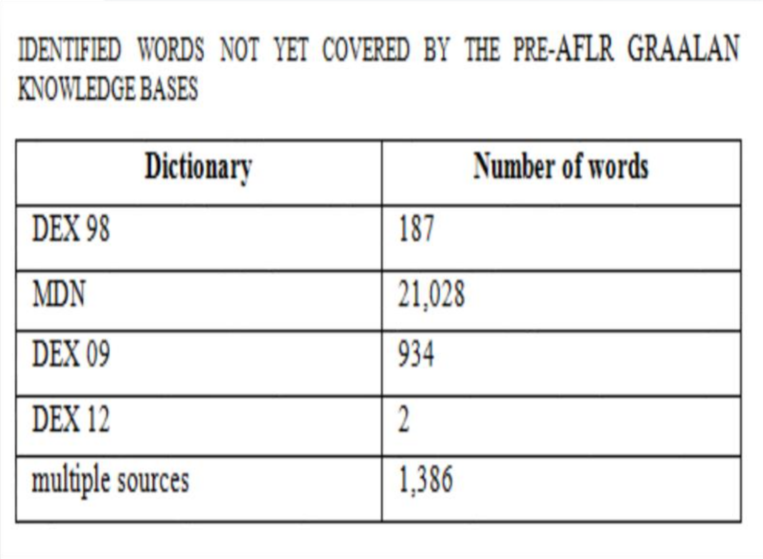
Challenges and Achievements

- The selected word candidates were introduced in the LKB by the members of the ILIR partner, using software instruments developed by SOFTWIN and the already defined GRAALAN phonetic, syllabification, and morphological rules (applied in order to automatically generate results from the linguistic input)
- The **newly added data** implied:
  - **Linguistic decisions regarding the phonetic transcription, the syllabification and/or the definitions of the words** that were not previously attested in the traditional, general dictionaries
  - **Refinement of the rules** applied by the software tools in order to automatically generate the desired output after receiving the linguistic data (e.g., **152 new phoneme-to-grapheme rules** –**in the form of GRAALAN groups** – **were ad**ded)
  - Tool updates

Challenges and Achievements

- **Digital online dictionaries inspected**
  - DEX 98, DEX 09, DEX 12, NODEX (2002), MDA (2002), MDN (2000)
- **Digital Corpora**
  - ZiareRom provided by the ILIR partner
  - SOFTWIN Digital Library
  - European Union documents (freely available on the official website)

- The processes of selecting and adding the new word-candidates, from corpora, to the linguistic database led to the **identification of words and meanings that are not registered in the current dictionaries**; it also implied collecting and building all the linguistic information requested by the GRAALAN system: verifying each word in context in order to detect its meaning(s) and to build its inflection forms
- **Some of the entries already existing in the database were enriched with new (not previously attested) meanings**



IDENTIFIED WORDS NOT YET COVERED BY THE PRE-AFLR GRAALAN KNOWLEDGE BASES

Dictionary	Number of words
DEX 98	187
MDN	21,028
DEX 09	934
DEX 12	2
multiple sources	1,386

Conclusions and Future Aims

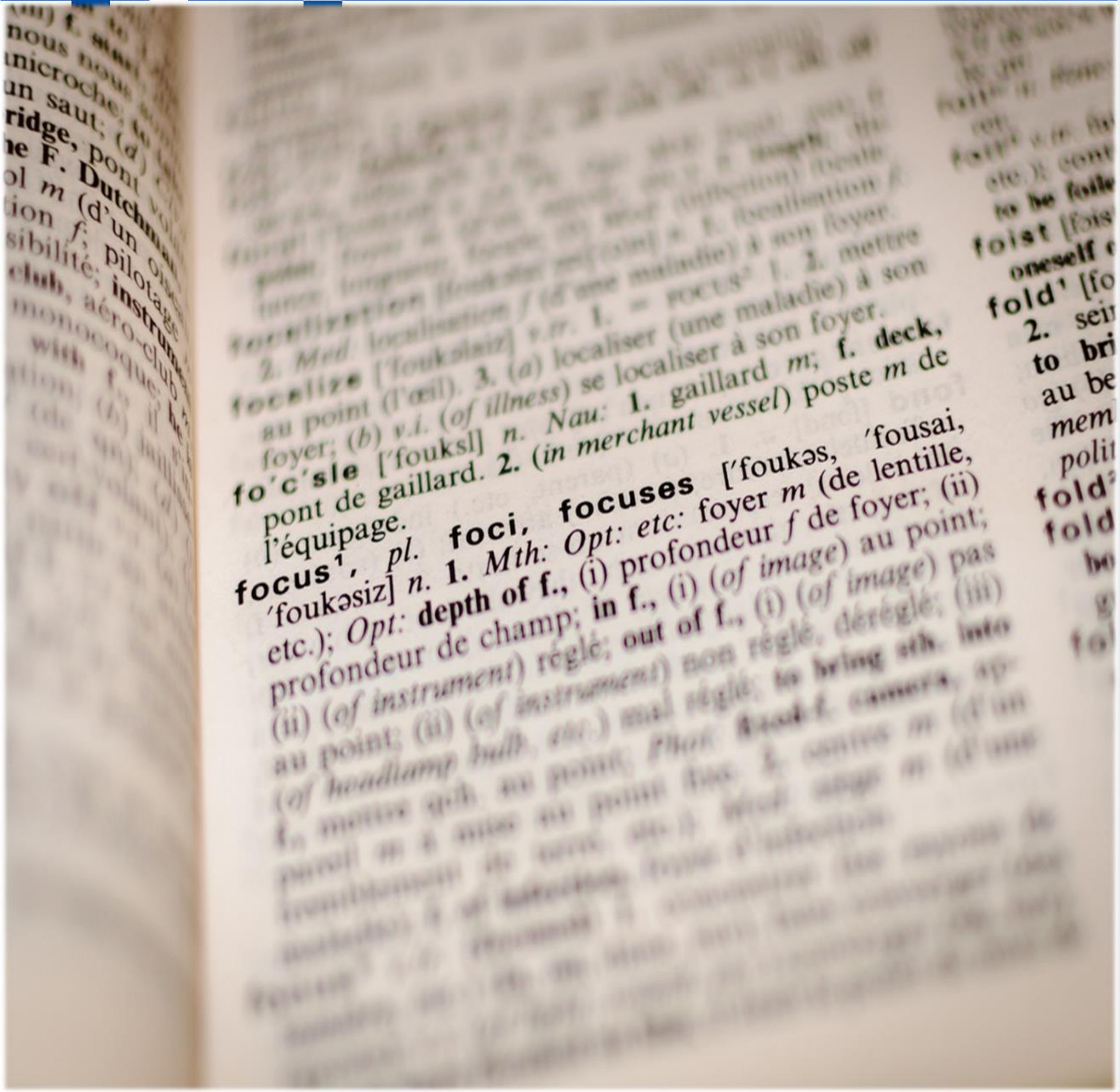
Results

- **Extended knowledge base** with more than 100,000 lemmas (covering **complex linguistic information such as phonetic transcriptions, syllabifications, glosses, morphology, semantic relations, inflection forms**) with a **high degree of accuracy** ensured by:
  - Our **rule-based method** applied for generating phonetic transcriptions (the GRAALAN-encoded rules are flexible and easily modifiable so as to reflect the linguists' decisions)
  - **Highly trained annotators**: the databases were filled exclusively by linguistic experts
  - **Strategies of verification at the end of each development stage**
- **Dictionaries that record (broad) phonetic transcriptions and syllable divisions/stress patterns using an internationally-approved standard**

Conclusions and Future Aims

Aims

- During AFLR we have mainly focused on introducing (single) words – we intend to enhance our database with **expressions/fixed phrases**
- The GRAALAN metalanguage is already designed in order to allow for MWE descriptions and tests are currently being performed



focalize ['foukəlaiz] v.tr. 1. = FOCUS² 1. 2. mettre au point (l'œil). 3. (a) localiser (une maladie) à son foyer; (b) v.i. (of illness) se localiser à son foyer.

fo'c'sle ['fouksl] n. Nau: 1. gaillard m; f. deck, pont de gaillard. 2. (in merchant vessel) poste m de l'équipage.

focus¹, pl. foci, focuses ['foukəs, 'fousai, 'foukəsiz] n. 1. Mth: Opt: etc: foyer m (de lentille, etc.); Opt: depth of f., (i) profondeur f de foyer; (ii) profondeur de champ; in f., (i) (of image) au point; (ii) (of instrument) réglé; out of f., (i) (of image) pas au point; (ii) (of instrument) non réglé, déréglé; (iii) (of headlamp bulb, etc.) mal réglé; to bring sth. into f.

Q&A