CIENSFO

Corpus of Non-Standard Spoken French Subordinated Interrogatives (Corpus d'Interrogatives Enchâssées Non-Standards du Français Oral)

Corpus content

This corpus contains transcriptions of spoken French sentences which exhibit non-standard subordinated interrogatives.

ex. ma façon de voir les choses c'est de faire le bilan de [pause] c'est quoi notre expertise

More precisely, five types of subordinated clauses are present:

1. Interrogatives **dependent on a noun**

- ex. moi je pouvais observer les différences de comment on était éduqués

2. Interrogatives dependent on a verb and **introduced by a preposition**

- ex. ça a pratiquement tout de suite reposé sur qu'est-ce qu'on va inventer comme femme [...]

3. Interrogatives being **verbal adjunct**

- ex. [...] ça peut être paronyme ou homonyme suivant comment vous le prononcez

4. Subordinated interrogatives using a **non-standard form** (ex. a marker unexpected in subordination)

- ex. avant ça je me pose jamais la question de est-ce que j'aime faire ça

5. Interrogatives used **in a nominal context**

- ex. avant de s'attaquer au à quoi ça sert commençons par le à quoi ça ne sert pas

TW: Some sources mention sensitive questions (sex, sexism, etc.). Thus, some sentences may use explicit words.

Corpus elaboration

This corpus is constituted with transcribed sentences personally observed by the authors. There are two kinds of sources.

1. heard in a **personnal conversations**: sentences uttered by someone else, sometimes the author being an interlocutor, sometimes not
2. from an **online** (often free) **material** (podcast, YouTube video, series, etc.)
3. saw written on an online forum or in a text message
4. sentences taken from the CEFC (see `cefc.tsv`)

Sentences of type 2. are provided with a complete description of their source (title, author, publisher, URL, time code), so that it is possible to check the transcription and obtain the actual prosody.

Sentences of type 1. were transcribed on the fly from a spontaneous conversation. There was no recording. Therefore, it is not possible to check the accuracy of the transcription nor to obtain the actual prosody. As a consequence, sentences of type 1. and 3. are little trustworthy. Please take that into account in your analyses.

Type 1. and 3. sentences were taken from persons of various age (but mostly French young adults) bewteen June 2022 and November 2023. Type 2 sentences are mostly extracted from materials put online between 2018 and November 2023 by persons of various age (but mostly French adults).

Transcription choices

Sentence segmentation is based on locutory units. A token `NAME` was substituted to proper names present in personal conversations.

Transcription is performed using standard lexical spelling forms. In particular, silent `e`'s are not removed (ex. `tu as` instead of `t'as` for `/ta/`). However, "missing words" are not added back (ex. `y a` instead of `y'a` or `il y a` for `/ja/`). Punctuation and capital letters starting sentences are not considered. The transcription follows the 1990 French spelling reform.

A `[pause]` symbol is added when there is a long enough pause between the predicate and the interrogative (only for type 2 sentences). Similarly, `euh`'s are not transcribed, except the ones close to the interrogative boarders.

ex. [...] un de mes objectifs c'est de partager avec vous mes réflexions sur [pause] comment vous pouvez vous créer votre propre mindset

A lot a extracts mentioned a list of interrogations following a first embedded interrogative. To avoid getting cumbersome lines, most of these additional interrogations were omitted.

Structure

The main document `ciensfo.json` is a json file. It contains a list of records. Each record contains fields:

- **id** (mandatory): unique sentence identifier
- **source** (mandatory): json description of the source of the sentence
- **time** (mandatory, except for type 1. and 2. sentences): time code of the online material `(hours:)minutes:seconds`
- **text** (mandatory): transcription
- **subtitles** (optional): official subtitles given by the publisher / author(s), they may differ from the transcription
- **modality** (optional): `written` or `spoken`, when absent, `spoken` is default
- **note** (optional): e.g. `ungrammatical, humoristic`, `maybe subrodinated exclamative`, `maybe free relative`, `maybe reported`
- **variant** (optional): variant of French, e.g. `Québec`, `Belgium`, when absent, default is European French

The source field has the following fields:

- **id** (mandatory): unique material identifier
- **title** (mandatory)
- **type** (mandatory): among:

  - type 1.: `conversation`, `scientific_conference`
  - type 2.: `online_podcast`, `series_epidose`, `radio_programme`, `recorded_speech`, `comedy_video`, `interview_video`, `newspaper_video`, `popularization_video`, `position_video`, `documentary_video`, `music`, `FAQ_video`, `tv_programme`
  - type 3.: `online_forum`, `comics`, `text_message`
- **date** (mandatory): online publication date, vector date format `[[year:month:day]]` (may be underspecified)
- **duration** (mandatory): `(hours:)minutes:seconds`
- **publisher** (optional)
- **catalog** (optional)
- **authors** (mandatory): the authors on the list may be identified by their given name, family name, literal name (ex. YouTube channel) or a mix of them
- **URL** (optional)
- **accessed** (mandatory, except for written): vector date format
- **page** (optional)
- **pages** (optional)
- **booktitle** (optional)
- **series** (optional)
- **volume** (optional)
- **ISSN** (optional)

Type 1. and 3. sentences only have source fields `id` and `type`.

Note that the person saying the extracted sentence may not always be one of the authors, but e.g. an interviewed person.

When some sentences are extracted from the same material, the source field of subsequent sentences may only contain the `id` field. Therefore, the couple "source" "id" + "time" (or just "source" "id" for type 1 sentences) constitute another possible unique sentence identifier.

Annotation

The file `classification_ciensfo.csv` contains, for each occurrence of a non-standard subordinated interrogatve, annotations about syntactic features. Theses labels have been added by hand.

Columns:

1. **sentence id**, if a sentence has several of such patterns, we add a dot and a second id. (e.g `14.1`, `14.2`). The id of CEFC sentences begins with a `c`
2. **sentence type**: 1, 2 or 3
3. **dependent on a noun**: if so, the field contains the lemma of the noun
4. **dependent on an adjective**: if so, the field contains the lemma of the adjective
5. **dependent on a verb (includes semi-fixed verbal expressions)**: if so, the field contains the lemma of the predicate (or `CONJ` if it is conjuncted with the previous interrogative in the same sentence)
6. **negated**: if dependent on a verb or attribute adjective, 1 if the predicate is negated or 0 is not
7. **adverbial adjunct clause**: if the interrogative is an adverbial modifier clause, the field contains the preposition(al locution) introducing it
8. **introducing preposition**: ("/" if no preposition)

9. **graft**: if the interrogative is a graft, the field contains the preceding word, typicaly a preposition or a determiner
10. **non-standard type**: if applicable:

- `qecq`: occurrence of *qu'est-ce que/qui*
- `ecq`: *est-ce que* instead of *si* or *WH + est-ce que* other than *qu'est-ce que/qui*
- `in-situ`: e.g *c'est quoi*
- `spp`: suffixed personal pronoun (aka. subject -verb inversion)

11. **WH**: interrogative word lemma
12. **WH 2**: second interrogative word lemma, if applicable
13. **marker**: interrogative marker
14. **marker**: additional morphosyntactic phenomenon which can hint at interrogativeness, e.g. *oui ou non, ou pas, ou non*
15. **class**: class according to [Coveney 2011]
16. **additional note**, e.g. `ungrammatical, humoristic`, `maybe subrodinated exclamative`, `maybe free relative`, `maybe reported`

Note: columns 3, 4, 5 and 7 may contain the token `CONJ` to indicate that the interrogative is conjuncted with the previous line, under the same governor.

Extended Coveney classification

The classification `type` (direct interrogatives only) is based on:

> Aidan Coveney. 2011. L'interrogation directe. Travaux de linguistique, 63(2):112–145. De Boeck Supérieur.

We extend it to account for infinitival interrogatives, subordinated interrogatives, nominal and elliptical interrogatives.

Note: Contrary to [Coveney 2011], `stats.py` considers expression *qu'est-ce que/qui* as an interrogative word, and not as `Q` + *est-ce que* pattern.

The list of categories is:

- yes-no interrogatives:
  - `ESV`: 'est-ce que', e.g. *Est-ce que les autres / ils sont partis ?*
  - `V-CL`: clitic inversion, e.g. *Sont-ils partis ?*
  - `GN V-CL`: complex inversion, e.g. *Les autres sont-ils partis ?*
  - `SV-ti`: '-ti' marker, *C'est-ti pas fini ?*
- constituent (fr. partielle):
  - `SVQ`: in situ, e.g. *Ils sont partis où ?*
  - `QSV`: fronting (fr. antéposition), e.g. *Où ils sont partis ?*
  - `QV-CL`: qu + clitic inversion, e.g. *Où sont-ils partis ?*
  - `Q GN V-CL`: qu + complex inversion, e.g. *Où les autres sont-ils partis ?*
  - `QV GN`: qu + stylistic inversion, e.g. *Où sont partis les autres ?*
  - `seQkSV`: cleft, e.g. *C'est où qu'ils sont partis ?*
  - `QESV`: qu + 'est-ce que', e.g. *Où est-ce qu'ils sont partis ?*
  - `QsekSV`: qu + cleft variant, e.g. *Où c'est qu'ils sont partis ?*

    - `QkSV`: qu + complementizer, e.g. *Où qu'ils sont partis ?*
    - `Q=S V`: subject qu, e.g. *Lesquels sont partis ?*
- hybrid (non-standard)
    - `QEV GN`: qu+ 'est-ce que' + stylistic inversion, e.g. *Avec qui est-ce que travaille nicole Dupont ?*
    - `Q=S V-CL`: subject qu + clitic inversion, e.g. *De ces fillettes, lesquelles sont-elles les tiennes ?*
    - `E GN V-CL`: 'est-ce que' + complex inversion, e.g. *Est-ce que demain les sauveteurs pourront-ils s'approcher des alpinistes en détresse ?*
    - `QE GN V-CL`: qu + 'est-ce que' + complex inversion e.g. *Qu'est-ce que le rédacteur de la rubrique des chats écrasés entend-il par un pachyderme ?*

Our extension includes:

- other case
    - `Q=S sekV`: subject qu + cleft *c'est qui* + verb, e.g. *Qui c'est qui diffuse ça ?*
- infinitival
    - `QVinf`: qu + infinitival verb, e.g. *Où partir ?*
    - `Vinf Q`: infinitival verb + in-situ qu, e.g. *Pour partir où ?*
    - `QsekVinf`: qu + cleft variant + infinitival verb, e.g. *Qu'est-ce que c'est qu'être une fille ?*
- multiple qu-words
    - `Q=S VQ`: double qu-interrogative with one qu subject, e.g. *Qui veut intervenir dans quoi ?*
    - `QSVQ`: double qu-interrogative, e.g. *Combien d'infanteries tu envoies sur quelle planète ?*
    - `QVinf QQ`: triple qu-interrogative + infinitival verb, e.g. *Qui inviter à quel endroit sur quel sujet ?*
- Nominal or elliptical
    - `Q GN`: qu + noun phrase, e.g. *Pourquoi Angiox ?*
    - `Qsek GN`: qu + cleft variant + noun phrase, e.g. *Qu'est-ce que c'est que l'énergie ?*
    - `Q`: elliptical qu (interrogative phrase alone), e.g. *Où ?*
- embedded yes-no
    - `si SV`: 'si', e.g. *Je sais s'ils sont partis.*
- other regional variants
    - `SV-tu`: '-tu' marker, *C'est-tu vraiment si pire que ça ?*

Corpus searches

The pattern created to search for interrogative adverbial modifier clauses use Grew (v. >= 1.14). Request files and found sentences are in the `searches` folder.

- `fib_comp_prep.req` used on the FIB (on the enriched version)
- `orfeo_prep_int.req` used on the CEFC corpus

The raw results can be found in the json files for the FIB and in the `CEFC_found` folder for the CEFC. The files with `_0` ending contain the pattern with the preposition immediately preceding the QU-word or morphosyntactic marking. The files with `_1` ending, where there is one word in between.

License

The data is collected under the Right to quote. It is distributed under the Creative Common CC-BY 4.0 license.

Publication

Please cite this publication to mention CIENSFO.

Richard, V. D. (2024). “selon coment vous vous positionnez” : Étude des circonstancielles à interrogative. to appear in 9e Congrès Mondial de Linguistique Française, Lausanne.

Credit

Author: Valentin D. Richard

Log

Counts are given excluing sentences from the CEFC. Word count is approximated by tokenizing on spaces and aportrophes.

- v1.0: Release version for the CMLF article (325 utterances, 7032 tokens, ~7513 words)