ДОДАТОК А

Приклади навчальних даних

Лістинг А.1 – Приклад структури датасету IWSLT 2017, англійська мова

```
...
<doc docid="531" genre="lectures">
<url>http://www.ted.com/talks/brian_cox_what_went_wrong_at_the_lhc</url>
<description>TED Talk Subtitles and Transcript: In this short talk from TED U 2009, Brian Cox shares what's new with the CERN supercollider. He covers the repairs now underway and what the future holds for the largest science experiment ever attempted.</description>
<keywords>talks, astronomy, energy, exploration, physics, science, technology</keywords>
<talkid>531</talkid>
<title>Brian Cox: What went wrong at the LHC</title>
<reviewer></reviewer>
<translator></translator>
<seg id="1"> Last year at TED I gave an introduction to the LHC.  </seg>
<seg id="2"> And I promised to come back and give you an update  on how that machine worked.  </seg>
<seg id="3"> So this is it. And for those of you that weren't there,  the LHC is the largest scientific experiment ever attempted --  27 kilometers in circumference.  </seg>
<seg id="4"> Its job is to recreate the conditions  that were present less than a billionth of a second after the universe began,  up to 600 million times a second.  </seg>
<seg id="5"> It's nothing if not ambitious.  </seg>
<seg id="6"> This is the machine below Geneva.  </seg>
<seg id="7"> We take the pictures of those mini-Big Bangs inside detectors.  </seg>
```

<seg id="8"> This is the one I work on. It's called the ATLAS detector -- 44 meters wide, 22 meters in diameter. </seg>

<seg id="9"> Spectacular picture here of ATLAS under construction so you can see the scale. </seg>

<seg id="10"> On the 10th of September last year we turned the machine on for the first time. </seg>

<seg id="11"> And this picture was taken by ATLAS. </seg>

<seg id="12"> It caused immense celebration in the control room. </seg>

<seg id="13"> It's a picture of the first beam particle going all the way around the LHC, colliding with a piece of the LHC deliberately, and showering particles into the detector. </seg>

<seg id="14"> In other words, when we saw that picture on September 10th we knew the machine worked, which is a great triumph. </seg>

<seg id="15"> I don't know whether this got the biggest cheer, or this, when someone went onto Google and saw the front page was like that. </seg>

<seg id="16"> It means we made cultural impact as well as scientific impact. </seg>

<seg id="17"> About a week later we had a problem with the machine, related actually to these bits of wire here -- these gold wires. </seg>

<seg id="18"> Those wires carry 13 thousand amps when the machine is working in full power. </seg>

<seg id="19"> Now the engineers amongst you will look at them and say, "No they don't. They're small wires." </seg>

<seg id="20"> They can do that because when they are very cold they are what's called superconducting wire. </seg>

<seg id="21"> So at minus 271 degrees, colder than the space between the stars, those wires can take that current. </seg>

<seg id="22"> In one of the joints between over 9,000 magnets in LHC, there was a manufacturing defect. </seg>

<seg id="23"> So the wire heated up slightly, and its 13,000 amps suddenly encountered electrical resistance. </seg>

<seg id="24"> This was the result. </seg>

<seg id="25"> Now that's more impressive when you consider those magnets weigh over 20 tons, and they moved about a foot. </seg>

<seg id="26"> So we damaged about 50 of the magnets. </seg>

<seg id="27"> We had to take them out, which we did. </seg>

<seg id="28"> We reconditioned them all, fixed them. </seg>

<seg id="29"> They're all on their way back underground now. </seg>

<seg id="30"> By the end of March the LHC will be intact again. </seg>

<seg id="31"> We will switch it on, and we expect to take data in June or July, and continue with our quest to find out what the building blocks of the universe are. </seg>

<seg id="32"> Now of course, in a way those accidents reignite the debate about the value of science and engineering at the edge. It's easy to refute. </seg>

<seg id="33"> I think that the fact that it's so difficult, the fact that we're overreaching, is the value of things like the LHC. </seg>

<seg id="34"> I will leave the final word to an English scientist, Humphrey Davy, who, I suspect, when defending his protege's useless experiments – his protege was Michael Faraday -- said this, "Nothing is so dangerous to the progress of the human mind than to assume that our views of science are ultimate, that there are no mysteries in nature, that our triumphs are complete, and that there are no new worlds to conquer." </seg>

<seg id="35"> Thank you. </seg>

</doc>

...

Лістинг А.2 – Приклад структури датасету IWSLT 2017, німецька мова

```
…
<doc docid="531" genre="lectures">

<url>http://www.ted.com/talks/brian_cox_what_went_wrong_at_the
_lhc</url>
      <description>TED Talk Subtitles and Transcript: In diesem
kurzen Vortrag der TED U 2009, präsentiert Brian Cox Neuigkeiten
vom CERN Teilchenbeschleuniger. Er spricht über die
gegenwärtigen Reparaturen und offenbart, was die Zukunft für das
größte wissenschaftliche Experiment der Geschichte
bereithält.</description>
      <keywords>talks, astronomy, energy, exploration, physics,
science, technology</keywords>
      <talkid>531</talkid>
      <title>Brian Cox: Was am LHC missglückte</title>
      <reviewer
href="http://www.ted.com/profiles/154067">Dominik
Weickgenannt</reviewer>
      <translator
href="http://www.ted.com/profiles/237190">Steffen
Lewenhardt</translator>
     <seg id="1">  Auf der letzten TED Konferenz gab ich eine
Einführung zum LHC.  </seg>
     <seg id="2">  Und ich versprach zurück zu kommen, um Ihnen
zu erklären wie die Maschine funktioniert.  </seg>
     <seg id="3">  Nun ist es also soweit. Und für alle jene die
damals nicht da waren, der LHC ist das größte wissenschaftliche
Experiment welches je angegangen wurde -- 27 Kilometer im
Durchmesser.  </seg>
     <seg id="4">  Seine Aufgabe ist es, die Bedingungen zu
erschaffen, welche weniger als eine Milliardstel Sekunde nach
Beginn des Universums existierten -- und das bis zu 600 Millionen
Mal innerhalb einer Sekunde.  </seg>
```

<seg id="5"> Es ist einfach nur atemberaubend ehrgeizig. </seg>

<seg id="6"> Dies ist die Maschine unterhalb von Genf. </seg>

<seg id="7"> Wir nehmen Bilder dieser Mini-Urknalle in Detektoren auf. </seg>

<seg id="8"> An diesem arbeite ich. Er nennt sich ATLAS Detektor -- 44 Meter in der Breite, 22 Meter im Durchmesser. </seg>

<seg id="9"> Hier ein spektakuläres Bild des ATLAS in der Konstruktion damit Sie die Größenverhältnisse sehen. </seg>

<seg id="10"> Am 10. September letzen Jahres lief die Maschine zum ersten Mal an. </seg>

<seg id="11"> Und dieses Bild wurde von ATLAS aufgenommen. </seg>

<seg id="12"> Es verursachte immensen Jubel im Kontrollraum. </seg>

<seg id="13"> Es ist ein Bild des ersten Strahlenpartikels welches die gesamte Strecke um den LHC zurücklegte, dann absichtlich mit einem Teil des LHC kollidierte, um einen Regen von Partikeln auf den Detektor prasseln zu lassen. </seg>

<seg id="14"> In anderen Worten, als wir dieses Bild am 10. September sahen, wussten wir, dass die Maschine funktioniert, was ein großer Triumph ist. </seg>

<seg id="15"> Ich weiss nicht, ob dies den größten Jubel erzeugte, oder dies, als jemand die Google Seite besuchte und die Startseite so vorfand. </seg>

<seg id="16"> Es bedeutet, dass wir einen kulturellen Einfluss neben dem wissenschaftlichen erreichten. </seg>

<seg id="17"> Ungefähr eine Woche später gab es ein Problem mit der Maschine welches mit diesen Drähten hier zu tun hatte -- diese goldenen Drähte hier </seg>

<seg id="18"> Diese Drähte leiten 13-tausend Ampere wenn die Maschine auf Hochleistung läuft. </seg>

<seg id="19"> Nun, die Ingenieure unter Ihnen werden sie betrachten und sagen, "Nein das tun sie nicht. Das sind kleine Drähte." </seg>

<seg id="20"> Sie können das leisten, weil wenn man sie sehr weit abkühlt, werden sie zu sogenannten Supraleitern. </seg>

<seg id="21"> Bei minus 271 Grad also, kälter als der Raum zwischen den Sternen, können diese Drähte die Spannung aushalten. </seg>

<seg id="22"> In einer der Verbindungen, zwischen über neuntausend Magneten im LHC, gab es einen Manufakturdefekt. </seg>

<seg id="23"> Dadurch erhitzen sich die Drähte geringfügig und 13-tausend Ampere begegneten plötzlich elektrischem Widerstand. </seg>

<seg id="24"> Dies war das Ergebnis. </seg>

<seg id="25"> Nun, das alles ist noch beeindruckender wenn Sie sich vorstellen, dass die Magneten über 20 Tonnen wiegen und sich um ca. 30 cm verschoben. </seg>

<seg id="26"> Wir beschädigten also ungefähr 50 der Magneten </seg>

<seg id="27"> und mussten sie entfernen, was wir auch taten. </seg>

<seg id="28"> Wir konditionierten sie neu, reparierten sie. </seg>

<seg id="29"> Sie sind nun alle wieder auf ihrem Weg zurück. </seg>

<seg id="30"> Gegen Ende März wird der LHC wieder funktionsfähig sein. </seg>

<seg id="31"> Wir werden ihn anschalten, und erwarten erste Daten im Juni oder Juli sammeln zu können, und setzen unsere Reise fort, um herauszufinden was die Bausteine des Universums sind. </seg>

<seg id="32"> Nun, natürlich, auf eine Art entfachen solche Unfälle erneut die Debatte um den Wert von Wissenschaft und

Ingenieurskunst an solchen Grenzen. Es ist leicht, so etwas abzulehnen. </seg>

<seg id="33"> Ich denke die Tatsache, dass es so schwer ist, die Tatsache, dass wir über unsere Grenzen hinaus greifen, bildet den Wert von Dingen wie dem LHC. </seg>

<seg id="34"> Ich werde meine abschließenden Worte einem englischen Wissenschaftler überlassen. Humphrey Davy, der, wie ich vermute, als er die nutzlosen Experimente seines Lehrlings verteidigte, sein Lehrling war Michael Faraday, folgendes sagte: "Nichts ist so gefährlich für die Entwicklung des menschlichen Geistes, als anzunehmen, dass unsere wissenschaftlichen Vorstellungen endgültig sind, dass es keine Mysterien in der Natur gibt, dass unsere Triumphe vollzählig sind, und dass es keine neuen Welten zu erobern gibt." </seg>

<seg id="35"> Vielen Dank. </seg>

</doc>

…

Таблиця А.1 – Приклад структури датасету IWSLT 2017, комбінований набір даних

Id	Англійська	Німецька
129-1	What I'm going to show you first, as quickly as I can, is some foundational work, some new technology that we brought to Microsoft as part of an acquisition almost exactly a year ago.	Was ich Ihnen zuerst, so schnell wie möglich, zeigen werde, ist die grundlegende Arbeit, eine neue Technologie, die wir Microsoft als Teil einer Übernahme vor genau einem Jahr, übergeben haben. Das ist Seadragon.

Продовження таблиці А.1

1	2	3
129-10	To prove to you that it's really text, and not an image, we can do something like so, to really show that this is a real representation of the text; it's not a picture.	Um zu beweisen, dass es sich wirklich um einen Text und nicht um ein Bild handelt, können wir etwas wie das tun, um wirklich zu zeigen, dass diese eine echte Repräsentation des Texts und kein Bild ist.
129-11	Maybe this is an artificial way to read an e-book.	Vielleicht ist dies eine künstliche Art, ein E-Buch zu lesen.
129-12	I wouldn't recommend it.	Ich würde es nicht empfehlen.
129-13	This is a more realistic case, an issue of The Guardian.	Dies ist ein realistischerer Fall. Dies ist eine Ausgabe von The Guardian.

Лістинг А.3 – Приклад структури датасету Europarl v7, англійська мова

```
<chp>
<sp>
<en> I will make sure that the President is fully apprised of that and that is so done .
<sp>
```

<en> Mr President , I do not know the extent of this tragedy , but if children are dead , I would perhaps suggest to the House that we observe a minute ' s silence .

<sp>

<en> We need to have more information about the full extent of the tragedy .

<en> EXPLANATIONS OF VOTE - Redondo Jiménez report (A5-0152 / 2000)

<sp>

<2en> Mr President , I voted for the Redondo report which regulates European Union interventions in the event of fires and atmospheric pollution .

<2en> This is certainly a positive measure .

<2en> I voted for the motion but I would prefer it if , in future , in implementing the proposal , we took into consideration the fact that many pensioners and elderly people live near forests and they are often the only people left in remote and countryside areas .

<2en> Fires are frequent in these areas and it would therefore be a good idea , or rather appropriate to use these people to help prevent fires .

<sp>

<2en> (PT) This report contains amendments to the current regulations on the protection of Community forests against atmospheric pollution and forest fires .

<2en> The amendments to the regulation on protecting forests against fires are particularly important for Portugal .

<2en> In fact , we must pay considerably greater attention to the Mediterranean woodlands that have been affected by forest fires , and this requires considerable financial resources for reforestation and for the necessary fire prevention measures .

<2en> For Portugal , where fires have destroyed large areas of forest , it is vital to increase appropriations in order to supplement national investment in building infrastructure , for example , and to help local authorities , private landowners '

associations or wasteland management bodies to buy fire prevention equipment .

<2en> This is why the increase to EUR 77 million in Community appropriations proposed by the rapporteur as opposed to the EUR 50 million proposed by the Commission is so important .

<2en> It is also important to create a Community information system on forest fires so that the situation at any given time is more widely known .

<2en> This will enable us to fight more effectively against forest fires and their causes .

...

Лістинг А.4 – Приклад структури датасету Europarl v7, німецька мова

<chp>

<sp>

Ich werde sicherstellen , daß die Präsidentin darüber umfassend in Kenntnis gesetzt wird und daß das so geschieht .

<sp>

Herr Präsident , ich weiß nicht , wie schwer der Unfall war , aber wenn dabei Kinder zu Tode gekommen sind , würde ich dem Parlament vorschlagen , vielleicht eine Schweigeminute einzulegen .

<sp>

Wir brauchen mehr Informationen über das ganze Ausmaß der Tragödie .

ERKLÄRUNGEN ZUR ABSTIMMUNG - Bericht Redondo Jiménez (A5-0152 / 2000)

<sp>

Herr Präsident , ich habe für den Bericht Redondo für Maßnahmen der Europäischen Union zum Schutz des Waldes gegen Brände und Umweltverschmutzung gestimmt .

Gewiß handelt es dabei um positive Verordnungen .

Ich habe zwar dafür gestimmt , doch würde ich mich freuen , wenn man künftig bei der Umsetzung dieser Vorschläge

berücksichtigt , daß in der Nähe der Wälder viele ältere Bürger und Rentner leben , die zuweilen als einzige in den ländlichen und abgelegenen Gebieten , wo es häufig zu Waldbränden kommt , geblieben sind .

Deshalb wäre es gut , ja sogar angebracht , diese Bürger bei der Brandbekämpfung einzusetzen .

<sp>

(PT) Dieser Bericht beinhaltet Änderungsvorschläge zu den geltenden Verordnungen über den Schutz des Waldes der Gemeinschaft gegen Luftverschmutzung und gegen Brände .

Für Portugal spielen besonders die Änderungsvorschläge eine wichtige Rolle , die die Verordnung zum Kampf gegen Waldbrände betreffen .

In der Tat ist es notwendig , dem mediterranen Wald stärker Beachtung zu schenken , der in erheblichem Maße von Waldbränden betroffen ist und für den deshalb mehr finanzielle Mittel bereit gestellt werden müssen , um die Wiederbevölkerung und die erforderlichen vorbeugenden Maßnahmen gegen Brände vorantreiben zu können .

Für Portugal , wo durch Brände bedeutende Waldgebiete zerstört wurden , ist eine Erhöhung der Mittel zur Ergänzung der nationalen Investitionen unerläßlich .

Zu diesen gehört auch der Bau von Infrastrukturen und die Anschaffung von Ausrüstungen zum Ausbau der Brandvorbeugung sowohl durch die Gemeinden als auch durch Verbände der privaten Eigentümer oder durch die Gremien , die das Brachland verwalten .

Deshalb ist überaus wichtig , die Gemeinschaftsmittel aufzustocken , bei denen die Berichterstatterin statt der 50 Mio. Euro im Vorschlag der Kommission einen Betrag von 77 Mio. Euro empfiehlt .

Ebenfalls notwendig ist die Einrichtung eines Waldbrand-Informationssystems der Gemeinschaft , um ein genaueres Bild von der aktuellen Situation erhalten und so Waldbrände und ihre Ursachen effektiver bekämpfen zu können .

...

Таблиця А.2 – Приклад структури датасету Europarl v7, комбінований датасет

ID репліки	Оригінал англійською	Текст англійською	Текст німецькою
0-0-0	true	I will make sure that the President is fully apprised of that and that is so done .	Ich werde sicherstellen , daß die Präsidentin darüber umfassend in Kenntnis gesetzt wird und daß das so geschieht .
0-1-0	true	Mr President , I do not know the extent of this tragedy , but if children are dead , I would perhaps suggest to the House that we observe a minute ' s silence .	Herr Präsident , ich weiß nicht , wie schwer der Unfall war , aber wenn dabei Kinder zu Tode gekommen sind , würde ich dem Parlament vorschlagen , vielleicht eine Schweigeminute einzulegen .
0-2-0	true	We need to have more information about the full extent of the tragedy .	Wir brauchen mehr Informationen über das ganze Ausmaß der Tragödie .
0-2-1	true	EXPLANATIONS OF VOTE – Redondo Jiménez report (A5-0152 / 2000)	ERKLÄRUNGEN ZUR ABSTIMMUNG – Bericht Redondo Jiménez (A5-0152 / 2000)
0-3-0	false	Mr President , I voted for the Redondo report which regulates European Union interventions in the event of fires and atmospheric pollution .	Herr Präsident , ich habe für den Bericht Redondo für Maßnahmen der Europäischen Union zum Schutz des Waldes gegen Brände und Umweltverschmutzung gestimmt .

Продовження таблиці А.2

1	2	3	4
0-3-1	false	This is certainly a positive measure .	Gewiß handelt es dabei um positive Verordnungen .
0-3-2	false	I voted for the motion but I would prefer it if , in future , in implementing the proposal , we took into consideration the fact that many pensioners and elderly people live near forests and they are often the only people left in remote and countryside areas .	Ich habe zwar dafür gestimmt , doch würde ich mich freuen , wenn man künftig bei der Umsetzung dieser Vorschläge berücksichtigt , daß in der Nähe der Wälder viele ältere Bürger und Rentner leben , die zuweilen als einzige in den ländlichen und abgelegenen Gebieten , wo es häufig zu Waldbränden kommt , geblieben sind .
0-3-3	false	Fires are frequent in these areas and it would therefore be a good idea , or rather appropriate to use these people to help prevent fires .	Deshalb wäre es gut , ja sogar angebracht , diese Bürger bei der Brandbekämpfung einzusetzen .
0-4-0	false	(PT) This report contains amendments to the current regulations on the protection of Community forests against atmospheric pollution and forest fires .	(PT) Dieser Bericht beinhaltet Änderungsvorschläge zu den geltenden Verordnungen über den Schutz des Waldes der Gemeinschaft gegen Luftverschmutzung und gegen Brände .

Продовження таблиці А.2

1	2	3	4
0-4-1	false	The amendments to the regulation on protecting forests against fires are particularly important for Portugal .	Für Portugal spielen besonders die Änderungsvorschläge eine wichtige Rolle , die die Verordnung zum Kampf gegen Waldbrände betreffen .
0-4-2	false	In fact , we must pay considerably greater attention to the Mediterranean woodlands that have been affected by forest fires , and this requires considerable financial resources for reforestation and for the necessary fire prevention measures .	In der Tat ist es notwendig , dem mediterranen Wald stärker Beachtung zu schenken , der in erheblichem Maße von Waldbränden betroffen ist und für den deshalb mehr finanzielle Mittel bereit gestellt werden müssen , um die Wiederbevölkerung und die erforderlichen vorbeugenden Maßnahmen gegen Brände vorantreiben zu können .

Продовження таблиці А.2

1	2	3	4
0-4-3	false	For Portugal , where fires have destroyed large areas of forest , it is vital to increase appropriations in order to supplement national investment in building infrastructure , for example , and to help local authorities , private landowners ' associations or wasteland management bodies to buy fire prevention equipment .	Für Portugal , wo durch Brände bedeutende Waldgebiete zerstört wurden , ist eine Erhöhung der Mittel zur Ergänzung der nationalen Investitionen unerläßlich .
0-4-4	false	This is why the increase to EUR 77 million in Community appropriations proposed by the rapporteur as opposed to the EUR 50 million proposed by the Commission is so important .	Zu diesen gehört auch der Bau von Infrastrukturen und die Anschaffung von Ausrüstungen zum Ausbau der Brandvorbeugung sowohl durch die Gemeinden als auch durch Verbände der privaten Eigentümer oder durch die Gremien , die das Brachland verwalten .
0-4-5	false	It is also important to create a Community information system on forest fires so that the situation at any given	Deshalb ist überaus wichtig , die Gemeinschaftsmittel aufzustocken , bei denen die Berichterstatterin statt der 50 Mio. Euro im Vorschlag der Kommission einen

		time is more widely known .	Betrag von 77 Mio. Euro empfiehlt .
0-4-6	false	This will enable us to fight more effectively against forest fires and their causes .	Ebenfalls notwendig ist die Einrichtung eines Waldbrand-Informationssystems der Gemeinschaft , um ein genaueres Bild von der aktuellen Situation erhalten und so Waldbrände und ihre Ursachen effektiver bekämpfen zu können .

ДОДАТОК Б

Лістинги елементів моделі

Лістинг Б.1 – Програмний код скрипта для обробки датасетів IWSLT 2017

```
     def main(args: Array[String]): Unit = {
  val langs = List("en", "de")
  for (fn <- fileNames) {processSets(fn, parseRawSet, langs)}
}

def processSets(fileName: String => String, parser: (String =>
String, String) => Map[String, String], langs: List[String]):
Unit = {
  val srcMap = parser(fileName, langs(0))

  val tgtMap = parser(fileName, langs(1))

  val writer = new
FileWriter(s"$baseFolder${fileName("combined")}.tsv")
  writer.write(s"id\t${langs(0)}\t${langs(1)}\n")
  for (k <- srcMap.keys) {
    writer.write(s"$k\t${srcMap(k)}\t${tgtMap(k)}\n")
    writer.flush()
  }
  writer.close()
}

def parseSet(fileName: String => String, lang: String):
Map[String, String] = {
  val src = XML.loadFile(s"$baseFolder${fileName(lang)}")
  (for {
    doc <- src \ "_" \ "doc"
    docId = (doc \ "@docid").text
    segment <- doc \ "seg"
    segmentId = (segment \ "@id").text
  } yield {
    (s"$docId-$segmentId", segment.text)
  }).toMap
}

def parseRawSet(fileName: String => String, lang: String):
Map[String, String] = {
  val src = XML.loadFile(s"$baseFolder${fileName(lang)}")
  (for {
    doc <- src \ "_" \ "doc"
    docId = (doc \ "@docid").text
    row <- doc.child.collect({case Text(x) =>
x}).flatMap(_.split("\n")).filter(_.trim.nonEmpty).zipWithInde
```

```
x
  } yield {
    (s"$docId-${row._2}", row._1)
  }).toMap
}
```

Лістинг Б.2 – Програмний код скрипта для обробки датасетів Europarl v7

```
       def parseData(folder: String): Unit = {
         for (mode <- modes) {
           println("processing mode " + mode)
           val langs = folder.replace("\\", "").split("-")
           val sets = langs.map(lang => {
             val src = Source.fromFile(fullFileName(lang, mode,
folder))
             val data = src.getLines().toList
             src.close()
             (lang,
               lang match {
                 case "English" => processEn(data)
                 case _ => processDefault(data, lang)
               }
             )
           })
           val langCol1 = sets(0)._2
           val langCol2 = sets(1)._2
           val writer = new FileWriter(datasetName(mode,
folder))
           for {
             row1 <- langCol1
             row2 <- langCol2
             if row1("id") == row2("id")
           } {
             val row = (row1.toList ++ row2.toList).toMap
             writer.write(s"${row("id")}\t${row("en-
original")}\t${row(langs(0))}\t${row(langs(1))}\n")
             writer.flush()
           }
           writer.close()
           println("finished processing mode " + mode)
         }
       }

       def processEn(data: List[String]): List[Map[String,
Any]] = {
         data.mkString
           .split("<chp>").filterNot(_ == "")
           .zipWithIndex
           .flatMap(convoPair => {
             val (convo, convoId) = convoPair
             convo
```

```
                .split("<sp>").filterNot(_ == "")
                .zipWithIndex
                .flatMap(speechPair => {
                  val (speech, speechId) = speechPair
                  val isOriginal = speech.contains("<en>")
                  speech.split(if (isOriginal) "<en>" else
"<2en>").filterNot(_ == "")
                    .zipWithIndex
                    .map(phrasePair => {
                      val (phrase, phraseId) = phrasePair
                      Map(
                        "id" -> s"$convoId-$speechId-$phraseId",
                        "English" -> phrase,
                        "en-original" -> isOriginal
                      )
                    })
                })
            }).toList
        }

        def processDefault(data: List[String], language:
String): List[Map[String, Any]] = {
          data
            .map(r => if (r.startsWith("<chp>") ||
r.startsWith("<sp>")) r else "<br> " + r)
            .mkString
            .split("<chp>").filterNot(_ == "")
            .zipWithIndex
            .flatMap(convoPair => {
              val (convo, convoId) = convoPair
              convo
                .split("<sp>").filterNot(_ == "")
                .zipWithIndex
                .flatMap(speechPair => {
                  val (speech, speechId) = speechPair
                  speech.split("<br>").filterNot(_ == "")
                    .zipWithIndex
                    .map(phrasePair => {
                      val (phrase, phraseId) = phrasePair
                      Map(
                        "id" -> s"$convoId-$speechId-$phraseId",
                        language -> phrase,
                      )
                    })
                })
            }).toList
        }

        def main(args: Array[String]): Unit = {
          parseData(frenchFolder)
        }
```

Б.3 – Алгоритми навчання моделі з кешем

```
    @tf.function
    def train_step_optimized(inp, targ, enc_hidden):
        return train_step(inp, targ, enc_hidden)

    def train_step(inp, targ, enc_hidden):
        loss = 0
        with tf.GradientTape() as tape:
            enc_output, enc_hidden = encoder(inp, enc_hidden)
            dec_hidden = enc_hidden
            dec_input =
tf.expand_dims([targ_lang.word_index['<start>']] * BATCH_SIZE,
1)
            # Teacher forcing - feeding the target as the next
input
            for t in range(1, targ.shape[1]):
                # passing enc_output to the decoder
                predictions, dec_hidden, _ =
decoder(dec_input, dec_hidden, enc_output)
                loss += loss_function(targ[:, t], predictions)
                # using teacher forcing
                dec_input = tf.expand_dims(targ[:, t], 1)
        batch_loss = (loss / int(targ.shape[1]))
        if decoder.cache_enabled:
            variables = decoder.cache.trainable_variables
        else:
            variables = encoder.trainable_variables +
decoder.trainable_variables
        gradients = tape.gradient(loss, variables)
        optimizer.apply_gradients(zip(gradients, variables))
        return batch_loss

    def train_plain():
        decoder.cache_enabled = False

        for epoch in range(EPOCHS):
            start = time.time()
            enc_hidden = encoder.initialize_hidden_state()
            total_loss = 0
            for (batch, (inp, targ)) in
enumerate(dataset.take(steps_per_epoch)):
                batch_loss = train_step_optimized(inp, targ,
enc_hidden)
                total_loss += batch_loss
                if batch % 100 == 0:
                    print(f'Epoch {epoch + 1} Batch {batch}
Loss {batch_loss.numpy():.4f}')
            if (epoch + 1) % 2 == 0:
                checkpoint.save(file_prefix=checkpoint_prefix)
            print(f'Epoch {epoch + 1} Loss {total_loss /
steps_per_epoch:.4f}')
```

```
            print(f'Time taken for 1 epoch {time.time() -
start:.2f} sec\n')

    def train_cache():
        decoder.cache_enabled = True
        for epoch in range(CACHE_EPOCHS):
            start = time.time()
            enc_hidden = encoder.initialize_hidden_state()
            total_loss = 0
            for (batch, (inp, targ)) in
enumerate(dataset.take(steps_per_epoch)):
                batch_loss = train_step(inp, targ, enc_hidden)
                total_loss += batch_loss
                if batch % 100 == 0:
                    print(f'Epoch {epoch + 1} Batch {batch}
Loss {batch_loss.numpy():.4f}')
            print(f'Epoch {epoch + 1} Loss {total_loss /
steps_per_epoch:.4f}')
            print(f'Time taken for 1 epoch {time.time() -
start:.2f} sec\n')
```

Лістинг Б.4 – Алгоритм навчання розмовної моделі

```
    @tf.function
    def train_step_optimized(inp, targ, enc_hidden):
        return train_step(inp, targ, enc_hidden)

    def train_step(inp, targ, enc_hidden):
        loss = 0
        with tf.GradientTape() as tape
            enc_output, enc_hidden = encoder(inp, enc_hidden)
            dec_hidden = enc_hidden
            if train_convo:
                o_src, o_tgt = prepare_src_tgt_contexts(inp,
targ, enc_hidden)
                dec_hidden = decoder.perform_init_dec(o_src,
o_tgt)
            dec_input =
tf.expand_dims([targ_lang_tokenizer.word_index['<start>']] *
BATCH_SIZE, 1)
            # Teacher forcing - feeding the target as the next
input
            for t in range(1, targ.shape[1]):
                # passing enc_output to the decoder
                predictions, dec_hidden, _ =
decoder(dec_input, dec_hidden, enc_output)
                loss += loss_function(targ[:, t], predictions)

                # using teacher forcing
                dec_input = tf.expand_dims(targ[:, t], 1)
        batch_loss = (loss / int(targ.shape[1]))
        if decoder.cache_enabled:
```

```
            variables = decoder.cache.trainable_variables
        else:
            variables = encoder.trainable_variables +
decoder.trainable_variables
        gradients = tape.gradient(loss, variables)
        optimizer.apply_gradients(zip(gradients, variables))
        return batch_loss

    def train_plain():
        for epoch in range(EPOCHS):
            start = time.time()
            enc_hidden = encoder.initialize_hidden_state()
            total_loss = 0
            for (batch, (inp, targ)) in
enumerate(linear_dataset.take(steps_per_epoch)):
                batch_loss = train_step_optimized(inp, targ,
enc_hidden)
                total_loss += batch_loss
                if batch % 100 == 0:
                    print(f'Epoch {epoch + 1} Batch {batch}
Loss {batch_loss.numpy():.4f}')
            linear_losses.append(total_loss / steps_per_epoch)
            print(f'Epoch {epoch + 1} Loss {total_loss /
steps_per_epoch:.4f}')
            print(f'Time taken for 1 epoch {time.time() -
start:.2f} sec\n')

    def train_convo():
        for epoch in range(EPOCHS):
            start = time.time()
            enc_hidden = encoder.initialize_hidden_state()
            total_loss = 0
            for (batch, (inp, targ)) in
enumerate(linear_dataset.take(steps_per_epoch)):
                batch_loss = train_step(inp, targ, enc_hidden)
                total_loss += batch_loss
                if batch % 100 == 0:
                    print(f'Epoch {epoch + 1} Batch {batch}
Loss {batch_loss.numpy():.4f}')
            linear_losses.append(total_loss / steps_per_epoch)
            print(f'Epoch {epoch + 1} Loss {total_loss /
steps_per_epoch:.4f}')
            print(f'Time taken for 1 epoch {time.time() -
start:.2f} sec\n')
```

ДОДАТОК В

Відомість кваліфікаційної роботи

Позначення	Найменування	Дод. відомості
	Текстові документи	
1.	Пояснювальна записка	91 с.
	Інші документи	
2.	Презентаційні матеріали	плакатів

.	.	Прізвище та ініціали.	Підп.	Дата			
Розробив		Волков Є. Д.			Дослідження методів нейронного машинного перекладу в реальному часі на прикладі листування багатонаціональної групи	Шифр групи	Код напр./спец.
Перевірив		Чала Л.Е.				СШІм-19-2	122
Н.контр.		Малєєва І.А.				ХНУРЕ кафедра ШІ	
Затв.		Філатов В.О.					